THE PROCEEDINGS OF
THE PHYSICAL SOCIETY

VOL. 59, PART 4 1 July 1947 No 334

SOME OBSERVATIONS OF THE MAXIMUM FREQUENCY OF RADIO COMMUNICATION OVER DISTANCES OF 1000 KM. AND 2500 KM.

BY W. J. G. BEYNON,
The National Physical Laboratory; now at University College, Swansea

MS. received 27 *September* 1946; *read* 15 *November* 1946

ABSTRACT. Measurements have been made on the magnitude of the discrepancies to be expected in the practical application of maximum usable frequency (M.U.F.) calculations. The maximum usable frequencies for radio transmission over distances of 1000 km. and 2500 km. have been deduced from continuous observations at sunrise and sunset of the relative field strength of signals received at Slough from short-wave broadcasting stations located near Berlin and Moscow respectively. It was found that the discrepancy between mean calculated and observed values of M.U.F. amounted to -3% for 1000 km. and to -11% for 2500 km. From a critical examination of the results it is concluded that a large proportion of these errors, particularly for the distance 2500 km., was due to inadequate knowledge of the ionosphere characteristics near the mid-point of the trajectory, and that the real errors of the mean calculated values probably do not exceed $\pm 2\%$ at 1000 km. and $\pm 3\%$ at 2500 km.

§ 1. INTRODUCTION

In practical radio-communication problems, an accurate knowledge of the maximum usable frequency which can be transmitted from one point to another is of major importance. Several methods of calculating this quantity from vertical-incidence ionospheric data are available (Smith, 1938; Millington, 1938; Appleton and Beynon, 1940), and curves of the maximum usable frequency (M.U.F. curves) now form part of the regular ionospheric data issued from observatories. Qualitative information obtained from the practical application of these curves indicates that there is often considerable discrepancy between these calculated values of the M.U.F. and the experimental values; in particular, it appears that frequencies considerably in excess of the values calculated from vertical-incidence data can often be satisfactorily transmitted from one point to another. In seeking for the explanation of this apparent divergence between theory and experiment it is instructive to note the simplifications and assumptions usually involved in these calculations. The following briefly summarizes the major sources of error involved in calculating and applying M.U.F. curves.

(*a*) The application of vertical-incidence ionospheric data to oblique-transmission problems usually depends directly or indirectly on two fundamental theorems due respectively to Breit and Tuve (1926) and to Martyn (1935). In the original form these theorems were stated only for the case of a plane earth and ionosphere, and for distances of transmission greater than about 600 km. Some modifications must be introduced to compensate for the effect of the curvature of the earth and ionosphere. An exact compensation would complicate the mathematical analysis too considerably, so that practical solutions are always subject to some degree of approximation.

(*b*) The calculations are normally concerned with ordinary-ray transmission, the effect of the earth's magnetic field being subsequently added as a comparatively small correction term. The magnitude of this correction will depend on many factors, such as the ratio of the frequency to the gyro-frequency, and the distance and direction of transmission.

(*c*) It is usually assumed that for that part of the trajectory which lies within the ionosphere there is no horizontal gradient of ionization.

(*d*) In applying M.U.F. curves to practical cases, the ionospheric characteristics at the mid-point of the trajectory have usually to be interpolated from the characteristics measured at a limited number of observing stations. Some part of any discrepancy between calculated and observed values of the M.U.F. may thus be the result of inadequate knowledge of the controlling ionospheric conditions.

A really accurate test of the theoretical analysis underlying the calculation of the M.U.F. will require oblique-incidence pulse experiments over long distances of transmission, together with simultaneous normal-incidence observations at one or more intermediate stations along the path. Unfortunately, under present circumstances, this is not possible, but some interesting practical information can be obtained from observations on signals received from distant short-wave broadcasting stations. The experiments described in this paper compare observed and calculated maximum usable frequencies over sender-receiver distances of 1000 km. and 2500 km. for a practical case in which normal-incidence ionospheric data are available only at one end of the transmission path. An estimate of the accuracy which is obtained over these distances should prove useful in predicting the accuracy which is normally to be expected in oblique transmission over greater distances.

§ 2. THEORETICAL CONSIDERATIONS

Further reference will now be made to three of the points noted above concerning the calculated values of the M.U.F.

(a) *Effect of the earth's magnetic field*

In all the calculations involved in the present experiments a simple approximation has been adopted. It is assumed that in east-west transmission of frequencies near 10 Mc./s. over a distance of 1000 km., the M.U.F. for the extraordinary component exceeds that for the ordinary component by 0·2 Mc./s. For the Moscow–Slough transmissions the corresponding figure is taken to be 0·3 Mc./s. These values represent the order of the separation between the maximum usable frequencies of the two magneto-ionic components deduced from a simplified theoretical consideration of the factors involved.

(b) *Controlling ionospheric conditions*

In the experiments described here, the senders are located to the east of the receiver, and vertical-incidence ionospheric data are available at the receiving site only. Now for transmission paths along a parallel of latitude, when no direct ionospheric data are available, it is customary to assume that the variation in ionospheric characteristics with longitude is the same as, or at least similar to, the diurnal variation which is observed at any fixed station at that latitude. This is probably a reasonably valid assumption when dealing with sender-receiver distances up to about 1000 km., but for longer distances it may be expected to become progressively less and less accurate. In the present case this assumption is initially made, and subsequently it is reconsidered in the light of the actual results.

(c) *Calculated values of the* M.U.F.

The experimental results have been considered in relation to the analysis for a parabolic type of layer described in papers by Appleton and Beynon (1940, 1942). The important points involved in calculating the M.U.F. by this method are briefly given below.

For a parabolic type of reflecting layer the vertical-incidence relation between equivalent height of reflection h' and frequency f can be represented by the equation

$$h' = h_0 + \frac{x \,.\, y_m}{2} \log_e \frac{1+x}{1-x},$$

where

$$x = f/f^0.$$

f^0 is the ordinary-ray critical frequency, and y_m and h_0 are constants of the layer which can readily be determined from the experimental vertical-incidence (h', f) curve. When f^0, y_m and h_0 are known, the maximum usable frequency for any distance of transmission can be read from a graph. It can readily be shown that the value of the calculated M.U.F. depends principally upon the magnitude of $(y_m + h_0)$, and only to a very much lesser extent on the individual values of y_m and h_0. The vertical-incidence critical frequency f^0 can usually be measured to within $\pm 0{\cdot}1$ Mc./s., and $(y_m + h_0)$ can be determined to within ± 10 km. If f^0, y_m and h_0 are parameters of the controlling part of the ionosphere, a single calculated value of the maximum usable frequency for distances of about 1000 and 2500 km. should then be accurate to within about $\pm 3\%$ and $\pm 4\%$ respectively.

§3. EXPERIMENTAL PROCEDURE

The first observations were made in November 1942 on the signal received at Slough (lat. 51° 30′ N., long. 0° 33′ W.) from the Zeesen high-frequency broadcasting stations presumed to be situated at Königswusterhausen (lat. 52° 18′ N., long. 13° 37′ E.). Some of the reasons which prompted a study of transmissions from these senders are given below.

(*a*) The transmission path lies approximately in an east-west direction.

(*b*) The distance of Zeesen from Slough (990 km.) is not too large, and this made it probable that the controlling ionospheric conditions at the mid-point of the trajectory could be related to ionospheric conditions at Slough.

(*c*) The frequencies of the Zeesen senders were such as to ensure that one or more observations of the M.U.F. could be made during each sunrise or sunset period.

(*d*) At this latitude, ionospheric conditions are particularly suitable at the mid-winter period for experiments of the kind to be described below. There is a maximum diurnal change in region-F_2 ionization and a minimum occurrence of abnormal or intense region-E ionization.

The success of these initial experiments over 1000 km. prompted the extension of the observations to signals received from the short-wave broadcasting station situated near Moscow (lat. 55° 45′ N., long. 37° 37′ E.). The distance Moscow–Slough (2500 km.) is still well within the limit for single-hop transmission by region F_2, and the frequency of transmission was such as to ensure that signals from this station regularly became critical during the winter sunrise period. Observations necessarily had to be confined to the sunrise period, since it was found that in the afternoon transmissions on this frequency ceased before the frequency became critical. Observations on the Moscow signals were commenced in January 1943 and continued during the two succeeding winter periods.

For the Moscow–Slough path, the controlling point in the ionosphere is located at lat. 54° 30′ N., long. 17° 30′ E. In the absence of direct normal-incidence ionospheric data it would thus seem appropriate to use mean data deduced from normal-incidence observations at Slough (lat. 51° 30′ N., long. 0° 33′ W.) and from Burghead (lat. 57° 42′ N., long. 3° 30′ W.). We shall therefore assume initially that the longitude variation in ionospheric characteristics near the parallel of lat. 54° 30′ N. follows the mean of the diurnal variations observed at Slough and Burghead.

Measurements of the maximum usable frequency were obtained from a continuous record of the field strength of the received signal during the sunrise and sunset periods. Accurate observations were made of the time at which the strong ray transmission gave place to weak scatter signals at sunset or *vice versa* at sunrise. Normally it was quite easy to ascertain this critical time to within a minute or so, and at this instant the frequency of the transmission under observation was also the M.U.F. for that particular distance and direction of transmission. These field-strength observations were obtained automatically in the form of ink records of the voltage variations in the diode detector circuit of a commercial communication receiver. Preliminary tests indicated that the receiver was particularly free from frequency drift and that the tuned frequency remained accurately constant for many hours at a time. Even so, a careful check was maintained on the tuning of the receiver during the critical periods of the measurements. The aerial system was not calibrated, so that absolute field-strength values were not known, but the records gave relative values of the signal strength, this being all that was required. From measurements with a standard-signal generator it is estimated that the field strength of the received signal generally changed by a factor of at least 20 to 1 during the complete change from ray signal to "scattered" signal or *vice versa*. Tracings of actual records showing the transition from one type of transmission to the other are shown in figures 1 (*a*) and 2 (*a*). During the critical periods the automatic record was supplemented by aural observations of the received signal. Throughout the experiments, a careful watch was maintained on ionospheric conditions at Slough. These vertical-incidence observations consisted in visual measurements of the critical frequency of region F_2 at 15-minute intervals with a photographic (h', f) record once every 30 minutes.

§4. EXPERIMENTAL RESULTS

(a) *Zeesen–Slough results*

These observations were made during the two periods 30 November 1942 to 12 December 1942 and 26 January 1943 to 12 February 1943 on the 9·65 Mc./s. and 11·86 Mc./s. Zeesen senders. During this period a very large number of transitions from the one type of transmission to the other were observed, but for reasons which will become clear from the discussion given below, only those observations which should be associated with the normal sunrise and sunset changes in region F_2 are considered. The results are summarized in table 1, together with the calculated values of the maximum frequency. The longitude difference between the mid-point of the trajectory and that of Slough corresponds to a time delay of 28 minutes, so that the calculated values given in the table are those deduced from normal-incidence observations at Slough 28 minutes after the entry or exit of the ray signal, as the case might be. During both periods of observation, forty reliable measurements were obtained of the transition, and of this number, eight which could not be related to the vertical-incidence observations made at Slough were rejected. The term "no correlation" indicates that the M.U.F. calculated from Slough data was either stationary or varying in the wrong sense at the appropriate time with respect to the observed transition.

(b) *Moscow–Slough results*

These observations were made during three winter periods: January–February 1943, December 1943 to February 1944, and December 1944 to February 1945. During this period a total of 80 reliable observations of the critical times were obtained. Typical samples of 10 results from each period are given in table 2, together with the calculated values of the M.U.F. It will be noted that the average results in all three groups of observations show a divergence of 9 to 11% between calculated and observed values.

§5. DISCUSSION OF RESULTS

(a) *Zeesen–Slough observations*

It has already been noted that in radio-communication problems it is only in very isolated cases that direct ionospheric measurements are available for the mid-point of a trajectory, and in the practical application of normal-incidence data to communication problems it is usually necessary to make some assumptions about the variation in ionospheric characteristics between the actual ionospheric stations. In east-west transmissions, such as those considered in this paper, we are particularly interested in the accuracy with which the variation with longitude can be deduced from the diurnal variation observed at a single fixed station, but few direct measurements have yet been made to examine this point. Schafer and Goodall (1939), however, have made some simultaneous critical-frequency measurements at Washington and Deal, U.S.A. These two stations are 300 km. apart, Deal being N.E. of Washington, and the published curves of these authors show that detailed fluctuations in the critical-frequency values are often repeated at Washington some time after occurring at Deal, but these results do not indicate accurately the time which elapses between the repetition of such fluctuations at the two stations. In the present experiments, a comparison between the field-strength records and the vertical-incidence ionosphere measurements at Slough yields some interesting information on this point.

Table 1. Observations at Slough on signals from Zeesen

Date	Freq. (Mc./s.)	Entry of ray signal	Exit of ray signal	Calculated M.U.F. (Mc./s.)	Difference calc.–obs. (Mc./s.)	Percentage difference	Estimated sender-receiver distance (km.), *vide* § 6
		(G.M.T.)					
Results for period 30 *November* 1942 *to* 12 *December* 1942							
30.11.42	11·86	0824		12·0	+0·1	+0·8	980
	11·86		1426	12·2	+0·3	+2·5	950
	9·65		1554	9·8	+0·15	+1·5	965
1.12.42	11·86	0921		11·8	−0·1	−0·8	1000
	11·86		1408	11·2	−0·7	−5·9	1080
	11·86		1422	11·0	−0·9	−7·5	1100
	9·65		1544	9·8	+0·15	+1·5	970
2.12.42	9·65	0750		9·05	−0·6	−6·2	1085
	11·86		1427	11·3	−0·6	−5·0	1065
3.12.42	11·86	0909		12·1	+0·2	+1·7	965
	11·86		1304	11·6	−0·3	−2·5	1030
	11·86		1341	No correlation			
	9·65		1505	10·05	+0·4	+4·1	930
4.12.42	11·86		1440	11·1	−0·8	−6·7	1090
5.12.42	9·65	0806		10·2	+0·55	+5·6	905
	11·86	0845		11·8	−0·1	−0·8	1000
7.12.42	9·65	0814		9·5	−0·15	−1·5	1010
	11·86	0840		11·6	−0·3	−2·5	1030
	11·86		1454	11·7	−0·2	−1·7	1015
9.12.42	9·65		1513	No correlation			
	11·86		1427	No correlation			
10.12.42	11·86	0919		11·7	−0·2	−1·7	1015
12.12.42	9·65	0852		10·1	+0·45	+4·6	920
	11·86	0946		11·4	−0·5	−4·2	1055
Results for period 26 *January* 1943 *to* 12 *February* 1943							
26.1.43	11·86		1328	10·1	−1·8	−15·1	1215
	9·65		1548	9·0	−0·65	− 6·7	1090
28.1.43	11·86	0846		10·3	−1·6	−13·4	1190
1.2.43	9·65	0820		No correlation			
	11·86	0931		11·0	−0·9	−7·5	1100
2.2.43	9·65	0801		No correlation			
	9·65	0819		8·4	−1·25	−13·0	1185
	9·65		1551	10·1	+0·45	+4·6	920
3.2.43	9·65	0840		No correlation			
5.2.43	9·65	0938		9·75	+0·1	+1·0	975
8.2.43	9·65	0759		9·4	−0·25	−2·6	1030
9.2.43	9·65	0752		9·6	−0·05	−0·5	1000
	11·86	0936		No correlation			
10.2.43	9·65	0754		8·65	−1·0	−10·3	1145
11.2.43	9·65	0733		8·2	−1·45	−15·0	1215
12.2.43	9·65	0812		No correlation			

Table 2

Samples of M.U.F. observations at Slough on signals from Moscow.
Frequency of transmission 10·44 Mc./s.

Date	Time of entry of ray signal (G.M.T.)	Calculated M.U.F. (mean of Slough and Burghead ionospheric data) (Mc./s.)	Percentage error (%)
Results for period January–February 1943			
18.1.43	0700	10·5	0
20.1.43	0643	9·3	−11
21.1.43	0718	8·8	−16
9.2.43	0550	8·3	−20
11.2.43	0549	8·6	−18
16.2.43	0521	8·5	−19
19.2.43	0544	9·3	−11
23.2.43	0518	8·1	−22
25.2.43	0546	10·5	0
27.2.43	0548	10·5	0
		Average error	−11·7
		Average error of 17 calculated values in this period	−11%
Results for period December 1943–*February* 1944			
20.12.43	0628	8·7	−17
22.12.43	0655	10·1	− 5
4. 1.44	0637	9·3	−11
13. 1.44	0641	9·3	−11
19. 1.44	0655	10·6	+ 1
21. 1.44	0633	8·8	−16
25. 1.44	0626	10·3	− 1
27. 1.44	0631	9·0	−14
1. 2.44	0623	10·3	− 1
10. 2.44	0630	8·4	−20
		Average error	− 9·3
		Average error of 20 calculated values in this period	− 9%
Results for period December 1944–*February* 1945			
11.12.44	0636	9·7	− 7
19.12.44	0651	9·6	− 8
6. 1.45	0630	9·5	− 9
16. 1.45	0631	8·9	−15
20. 1.45	0616	10·2	− 2
25. 1.45	0612	9·3	−11
30. 1.45	0645	9·2	−12
3. 2.45	0615	9·3	−11
10. 2.45	0607	8·4	−20
16. 2.45	0558	9·9	− 5
		Average error	−10
		Average error for 43 calculated values in this period	−11%
		Average error for 80 calculated values in all three periods	−11%

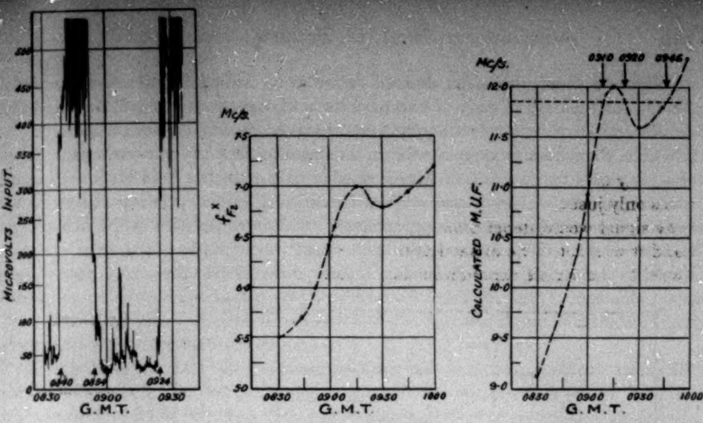
In some cases it is clear that the longitude variation in the ionosphere characteristics corresponds very closely with the local time variation observed at a fixed station, and in many other cases it is equally clear that there may be differences which are significant in relation to experiments of this kind. Figures 1 (*a*) and 1 (*b*) show an example of the detailed correspondence between the field-strength measurements on the Zeesen transmissions and the ionosphere characteristics observed at Slough. During the morning of 7 December 1942 the ray signal from Zeesen on a frequency of 11·86 Mc./s. was observed to come up sharply at 0840 G.M.T. The ray signal disappeared with almost equal abruptness at 0854 G.M.T. and reappeared again at 0924 G.M.T. The vertical-incidence critical frequency observed at Slough over this period is shown in figure 1 (*b*) and the calculated maximum usable frequency for a distance of 990 km. is shown in figure 1 (*c*). Comparing figures 1 (*a*) and 1 (*c*), it is clear that the oblique phenomena can be closely correlated with the ionosphere measurements made at the Slough end of the transmission path. The time delays between the actual entry, exit and re-entry of the ray signal and the times calculated from Slough data are 30, 26 and 22 minutes respectively. Assuming that local time and longitude variations in ionosphere characteristics are equivalent, the time delay corresponding to one-half the total transmission distance should be 28 minutes, so that in this particular example the observed time differences are quite near the expected value, but we shall see that this is not always the case. In the course of these experiments, comparatively small local irregularities in the ionosphere often caused the oblique signal to appear and disappear for short periods. On some days, particularly during the second group of observations, the signal often came up and disappeared again six or seven times during a period of a few hours. This is well illustrated in figure 2 (*a*), which shows the field-strength record of the 11·86 Mc./s. signal on 28 January 1943. The M.U.F. curve calculated from Slough data is shown in figure 2 (*b*), and, for convenience, the times at which the ray signal was present are also indicated in this figure. It will be noted that during the period of this record the ray signal appeared and disappeared no fewer than eight times. Most of these transitions can be correlated with maxima in the calculated M.U.F. values, but the time delays are clearly much larger than the 28 minutes which are normally assumed for this path length. A large proportion of the discrepancy between the calculated and observed M.U.F. for the sunrise period of this particular day (see table 1) can be ascribed to the fact that the time delay was of the order of 60 minutes rather than the assumed value of 28 minutes.

In the light of this discussion we now reconsider the observations given in table 1. The results for the period 26 January 1943 to 12 February 1943 differ from those for the earlier period in three main points:—

(*a*) The upper frequency (11·86 Mc./s.) was only received on very few occasions, and most of the observations were thus of necessity confined to the frequency 9·65 Mc./s.

(*b*) The discrepancy between calculated and observed values in this group of observations is considerably greater than that observed in the earlier period.

(*c*) There is an increased number of cases of "no correlation" with vertical-incidence Slough data.



(*a*) Recorded signal from Zeesen (11·86 Mc./s.).

(*b*) Vertical incidence measurements at Slough.

(*c*) Calculated oblique-incidence M.U.F.

Figure 1. Observations made at Slough, 7 December 1942.

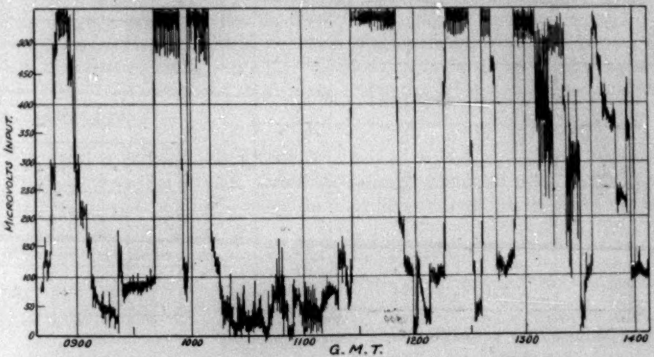


Figure 2 (*a*). Signal from Zeesen (11·86 Mc./s.) recorded at Slough, 28 January 1943.

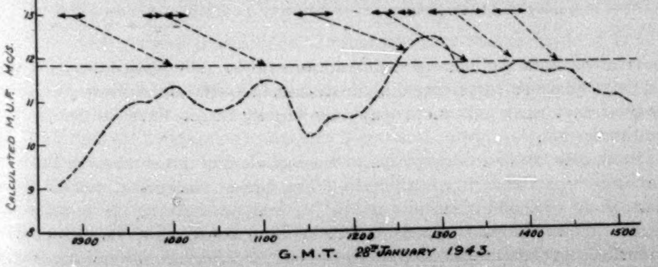


Figure 2 (*b*). Correlation of maxima in M.U.F. curve with observed 11·86 Mc./s. ray signal. Periods at which ray signal was observed are indicated thus ⟷

Figure 3 shows the mean diurnal variation of critical frequency measured at Slough for the days in each of two periods of observation. It will be noted that smaller values of critical frequency were observed during the second period. At this time, the critical frequency seldom increased sufficiently to permit observation of a ray signal from the 11·86 Mc./s. sender, and often the 9·65 Mc./s. ray signal was only just in. During such critical conditions, the reception or otherwise of a ray signal would depend on comparatively small irregularities in the ionosphere, and it was not to be expected that such small variations in ionosphere structure would be a simple function of solar angle. Now the M.U.F. calculations given here really assume that ionospheric conditions are a simple function of solar angle, and it is thus not surprising that these calculations should show the greatest error on days when the M.U.F. was only slightly larger than the frequency of communication.

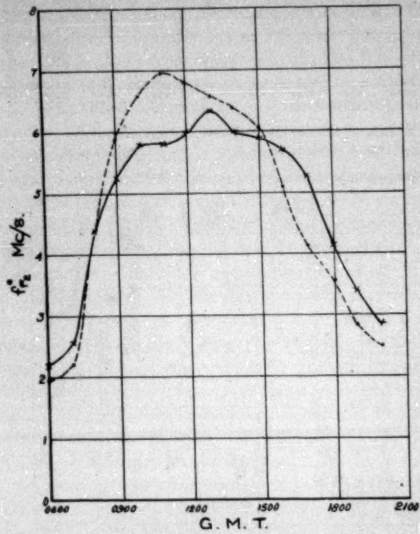


Figure 3. Mean values of $f^0_{F_2}$ at Slough.

30 Nov. 1942 to 12 Dec. 1942 •— — — —•

26 Jan. 1943 to 12 Feb. 1943 × ——— ×

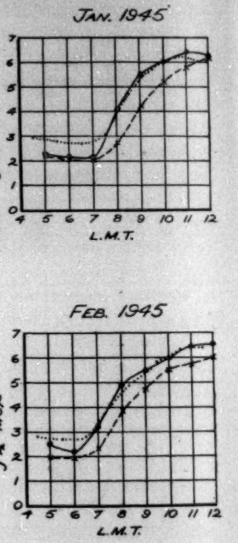


Figure 4. Monthly mean values of $f^0_{F_1}$.

×— — —× Burghead.

•.........• Moscow.

o———o Slough.

From many comparisons similar to those given above, it can be concluded that the time delay between corresponding ionosphere changes at points 500 km. apart, along a parallel of latitude at 51½° N., may be anything up to one hour or longer. The assumption of a fixed delay of 28 minutes, which has been made in the earlier part of this paper, is only valid when ionospheric conditions are changing quite rapidly, as is generally the case over the winter sunrise or sunset periods.

The experimental results for the Zeesen–Slough path may be summarized as follows.

Of the twenty-one values given in the first part of table 1, eight show an average positive error of 2·8% and thirteen an average negative error of 3·6%. It would appear that negative errors were slightly more probable than positive errors, but many of the errors are small and the relative numbers may not be significant. The arithmetic mean of these twenty-one calculated values differs from the observed value by −1·2% (probable error ±0·6%). This group of results suggests that for this distance and direction of propagation the systematic discrepancy between calculated and observed values of the M.U.F. does not exceed 2%. It will be noted that in nine cases the discrepancy between calculated and observed values exceeds the estimate of ±3% given in §2(c).

In the second group of observations there are only two cases in which the calculated M.U.F. exceeded that observed. These two positive errors have an average value of 2·8%, whereas the nine remaining negative errors have an average value of 9·3%. The arithmetic mean of the eleven calculated values is in error by −7·1%.

Considering all thirty-two observations together, we find that in ten cases the calculated M.U.F. exceeded that observed, the average discrepancy being 2·8%. In twenty-two cases the calculated M.U.F. was smaller than that observed, the average discrepancy being −6·0%. The average error (positive and negative) is about 5%. The arithmetic mean of the thirty-two calculated values is less than that observed by 3·2%.

(b) *Discussion of Moscow–Slough results*

Table 2 shows that in each period the average calculated value of the M.U.F. for the Moscow–Slough path, based on mean ionospheric data from Slough and Burghead, is some 11% too small.

We may note three possible causes for this discrepancy:

(i) For this distance of transmission the theoretical analysis may actually be in error by this amount.

(ii) At the critical sunrise period there might be appreciable lateral deviation of the incoming signal from the great-circle path.

(iii) The error may arise partly or entirely from incorrect assumptions about ionospheric conditions at the mid-point of the oblique path.

In considering these possibilities, it must be noted that an error of 11% is several times the magnitude of the error which we would have anticipated from the results for a distance of 1000 km., and would represent an extremely rapid rise in the divergence between theory and experiment when the distance is increased from 1000 km. to 2500 km. From the theoretical relationship between M.U.F. factor and distance for a "parabolic" reflecting layer, we should not expect the error at 2500 km. to be more than 1 or 2% larger than that observed for a distance of 1000 km.

The possibility of appreciable error due to abnormal lateral deviation of the signal at sunrise was investigated by taking bearing observations over the critical period on an Adcock direction finder. Tests on a sample of eight days, on each of which the calculated M.U.F. was much smaller than that observed, indicated conclusively that at the time of entry the ray signal was not deviated significantly from the true bearing value, and it would appear that in these particular experiments, this possibility is not a serious disturbing factor.

In considering the third possibility, we first note the results of a sample of observations on Moscow signals received at Burghead. The distance Moscow–Burghead (2460 km.) is only 2% less than the distance Moscow–Slough, so that the theoretical analysis should be equally accurate over the two paths. On the other hand the latitude of the mid-point of the Moscow–Burghead trajectory is practically equal to that of Burghead, and it should thus be possible to deduce ionospheric conditions at the mid-point from normal-incidence observations at Burghead alone. For measurements made at Burghead, there is thus reason to expect better agreement between theory and experiment than was noted in the case of the observations made at Slough. A group of ten such observations was made at Burghead in December 1943 and January 1944. In this case it was found that the mean calculated M.U.F. was 23% smaller than that observed. This very large error, under conditions in which we should have expected an error substantially smaller than 11%, suggests at once that in these experiments it is invalid to assume close correspondence between the local time variation and the longitude variation in ionosphere characteristics. It is also relevant to note that the error for these measurements at Burghead is almost exactly twice the magnitude of the error noted for the results given in table 2.

During the period of these Burghead observations a careful watch was also made at Slough on the time of entry of the Moscow ray signal. It was found that the mean time difference between the entry of the ray signal at Slough and Burghead was exactly equal to the time difference between sunrise at the two sites. Since the distances of Slough and Burghead from Moscow are practically equal, it could be inferred from this observation that, at the time of the experiments, ionosphere conditions governing the maximum usable frequencies for these two trajectories must have been identical.

The third point of interest is that for the 80 observations summarized in table 2, the mean M.U.F. calculated on Slough data alone is 10·63 Mc./s. (probable error ± 0·1 Mc./s.) and is thus within 3% of the correct value.

A little consideration of the three results given above will show that all three are quantitatively consistent with the conclusion that at the time of these experiments, ionospheric conditions governing the value of the M.U.F. at the mid-point of the Moscow–Slough trajectory, were very similar to those observed at Slough rather than to conditions midway between Slough and Burghead. Since these experiments were completed, substantial support for the correctness of this conclusion has come in the form of direct normal-incidence critical-frequency data from the ionospheric station recently in operation at Moscow. Figure 4 shows the diurnal variation in monthly mean values of $f^0_{F_2}$ measured at Slough, Burghead and Moscow for January and February 1945. Data from Moscow for December 1944 are, unfortunately, not available, but it is clear from the curves for January and February that over the period with which these experiments are concerned the monthly mean values of $f^0_{F_2}$ observed at Moscow are practically identical with the values observed at Slough.

Further confirmation of this type of variation in $f^0_{F_2}$ with longitude is provided when we examine critical-frequency data from other stations near this latitude but located further east. Thus if we consider data from Sverdlovsk (55° 50′ N., 60° 37′ E.) we find the winter values of $f^0_{F_2}$ at this station to be even higher than those observed at Moscow or at Slough. This marked longitude variation in critical frequency forms the subject of recent papers by Kessenikh and Bulatov (1944) and by Appleton (1946). It thus seems reasonable to assume that at the time of these experiments, near the mid-point of the Moscow–Slough path, conditions closely approximated to those given either at Moscow or Slough and, subject to this assumption, the mean observed value of the M.U.F. for 80 observations agrees with the mean calculated value to within 3%.

§6. ESTIMATION OF THE DISTANCE BETWEEN SENDER AND RECEIVER

In a paper by Appleton and Beynon (1942) a graphical representation has been given of the relation between M.U.F. factor and distance of transmission for a wide variety of ionospheric conditions. (The "M.U.F. factor" is the ratio of the M.U.F. to the normal-incidence critical frequency.) Hence if the M.U.F. factor be measured, it is a simple matter, from the curves, to estimate the distance between sender and receiver. At first it might appear that if the distance from the sender to receiver is unknown, then it will be impossible to measure the M.U.F. factor, since this implies a knowledge of ionospheric conditions at the mid-point of the trajectory. In the case of an east-west transmission path, however, we can overcome this difficulty by again assuming that the variation of ionosphere characteristics along the transmission path corresponds closely to the diurnal variation observed at vertical incidence at one end of the path. It is clear that these estimates of distance will be subject to similar errors as were the calculated values of the M.U.F. The actual estimates of sender-receiver distance for the Zeesen–Slough transmission path are given in the last column of table 1. For the whole group of thirty-two observations the mean calculated distance is 1036 km. The actual distance from Slough to Zeesen is 990 km., so that the actual error of the arithmetic mean is +46 km. or +4·6%. It is likely that if more accurate ionospheric data were available for the mid-point of the trajectory, then the accuracy of observations of sender-receiver distance made in this way would be considerably improved. If we consider the first group of results only, the mean calculated distance is 1005 km., so that the actual error in this case is only 15 km. or +1·5%.

It may be noted that for ranges beyond about 1000 km., and for a given accuracy in estimating the maximum usable frequency factor, there is theoretically a steady deterioration with increasing distance in the accuracy of a single estimate of sender-receiver distance. Thus an error of 2% in the M.U.F. factor for the Moscow–Slough path corresponds to a distance error of about 100 km.

§7. CONCLUSIONS

One of the objects of these experiments was to investigate the magnitude of the errors to be expected in a practical application of maximum usable frequency calculations. In the present experiments, full normal-incidence ionospheric data were available for one end of the transmission path and the transmitters were located in a direction approximately east of the receiving site. The conditions were thus rather more favourable than can be expected in the general application of such calculations. Nevertheless the results indicate that even for a transmission path of 1000 km., and with full normal-incidence ionospheric data for one end of the path, the divergence between individual calculated and observed values may occasionally amount to 15%. Results for the longer path of 2500 km. show differences of up to 25% between calculated and observed values in individual cases. In the case of the experiments over 1000 km., the fact that positive and negative errors were equally frequent suggests that there is no serious error in the analysis underlying the M.U.F. calculation. In the case of the measurements over 2500 km., the mean of 80 calculated values was 11% less than the observed value. The calculated value exceeded the observed value in seven cases only. There is evidence, however, that this discrepancy is almost entirely due to incorrect assumptions about ionospheric conditions near the mid-point of the trajectory. If such data are available it seems likely that the discrepancy between calculated and observed values of the maximum usable frequency for a distance of 2500 km. would not exceed 3%.

§8. ACKNOWLEDGMENTS

The results for the Zeesen transmissions described above were originally included in a confidential paper communicated in April 1943 to the Radio Research Board. The work was carried out as part of the programme of the Radio Research Board, and this paper is published by permission of the Department of Scientific and Industrial Research.

REFERENCES

Appleton, E. V., 1946. *Nature, Lond.*, **157**, 691.
Appleton, E. V. and Beynon, W. G., 1940. *Proc. Phys. Soc.*, **52**, 518.
Appleton, E. V. and Beynon, W. G. Radio Research Board, June 1942. (In course of publication.)
Breit, G. and Tuve, M. A., 1926. *Phys. Rev.*, II, **28**, 554.
Kessenikh, V. N. and Bulatov, H. D., 1944. *C.R. (Doklady) Acad. Sci. U.R.S.S.*, **45**, No. 6, 234.
Martyn, D. F., 1935. *Proc. Phys. Soc.*, **47**, 323.
Millington, G., 1938. *Proc. Phys. Soc.*, **50**, 801.
Schafer, J. P. and Goodall, W. M., 1939. *Terr. Mag. Atmos. Elect.*, **44**, 205.
Smith, N., 1938. *J. Res. Nat. Bur. Stds., Wash.*, **20**, 683.

DISCUSSION

on papers by:

(i) Sir E. Appleton and W. J. G. Beynon, F.R.S.: "The application of ionospheric data to radio communication problems" (*Proc. Phys. Soc.*, **59**, 58 (1947)).

(ii) W. J. G. Beynon: "Oblique radio transmission in the ionosphere and the Lorentz polarization term" (*Proc. Phys. Soc.*, **59**, 97 (1947)).

(iii) W. J. G. Beynon: "Some observations of the maximum frequency of radio communication over distances of 1000 km. and 2500 km." (*Proc. Phys. Soc.* **59**, 521 (1947)).

Mr. R. Dehn. With reference to the agreement of observations of overhead conditions of the ionosphere with those obtained from records of signals received and reflected at a distance equivalent to some 25 or 30 min. in time, would these conditions of the ionosphere be preserved for a longer period of time? That is, would the agreement be observable over greater distances?

Mr. F. A. Kitchen. Could prediction of maximum usable frequencies be made as much as twelve months ahead, following the correlation of distance factors with sunspot variation?

Mr. R. Naismith. A forecast of maximum usable frequencies for a period of twelve months ahead involves two quite separate requirements; an estimate of the vertical incidence conditions twelve months ahead; and an estimate of the maximum usable frequency from given vertical incidence conditions.

The authors have succeeded in supplying a very satisfactory solution to the latter requirement and have shown that the resultant error can be less than 3%.

An illustration of the comparative accuracy of the former may be given from current data. In any estimate of the vertical incidence critical frequency, account must be taken of the variation in the solar cycle among other factors. It is well known that the ionization in region F_2 varies in sympathy with the solar cycle when average values are coinsidered It may be assumed, therefore, that the ionization over the world would vary fairly uniformly due to this cause. Noon values of ionization for region F_2 in different parts of the world were compared for September 1945 and September 1946 (the latest month for which data are available), and it was found that whereas a 10% increase occurred in one part of the world, a 90% increase occurred in another. On present knowledge this large variation was quite unpredictable, and illustrates one of the difficulties in producing a forecast of ionospheric conditions so far ahead.

Dr. F. T. Farmer. I would like to know whether the calculations described take into account the earth's magnetic field for the ordinary wave, or whether it is only allowed for in calculating oblique frequencies for the extraordinary wave.

Author's reply. In reply to Mr. Dehn, it is possible that certain forms of ionospheric variation would persist for longer periods. Thus the detailed fluctuation in the sunrise change shown in figure 1 (p. 529) might occur at widely separated stations. Clear evidence of such correlation was noted only in the Zeesen–Slough experiments. In the case of the Moscow–Slough observations the repetition of detailed ionospheric changes was not obvious, but it is possible that a careful examination of the records would yield further information on this matter.

Concerning Mr. Kitchen's point, I think that Mr. Naismith has largely provided the answer. The main purpose of these papers was not to discuss prediction, but I may say that some predictions of the maximum usable frequency have been made for six months ahead, and, subject to a proportionate decrease in accuracy, an extension to a period of twelve months could, no doubt, be made. However, Mr. Naismith has already pointed out some of the difficulties involved.

In reply to Dr. Farmer, the effect of the magnetic field of the earth was not included in the calculations relating to the ordinary wave.

MEAN FREE PATH OF SOUND IN AN AUDITORIUM

By A. E. BATE and M. E. PILLOW,
Northern Polytechnic, London

MS. received 11 *October* 1946

ABSTRACT. A brief review of the subject is followed by proofs that the mean free path of sound in an enclosure is equal to 4 (Volume/Surface Area) for rectangular, spherical, and cylindrical rooms of any dimensions.

§1. INTRODUCTION

In some methods (Eyring, 1930) of calculating the reverberation time of sound in an auditorium in terms of the dimensions of the room and the absorptive properties of the walls, it is necessary to make use of a formula for the mean free path of sound, i.e. the mean distance travelled by all "rays" of sound between successive impacts with the walls.

Jaeger (1911), applying a method similar to that used in the kinetic theory of gases, obtained for this mean free path the value $4V/S$, where V represents the volume and S the internal surface area of the room. Jaeger's method shows that the value should be independent of the shape of the room and the position of the source, if uniform distribution of the sound energy is assumed: that is, if sound is travelling with equal intensity in all directions, through all points in the enclosure, at any instant considered.

Schuster and Waetzmann (1929) calculated the mean free path for rooms of certain simple shapes. Their values, which Eyring (1930) obtained at almost the same time by more approximate methods, are:

Cube	M.F.P. $= 2\sqrt{3}\,V/S = 3{\cdot}5\,V/S$,
Cylinder (height = diameter)	M.F.P. $= 3\sqrt{2}\,V/S = 4{\cdot}2\,V/S$,
Sphere	M.F.P. = diameter $= 6\,V/S$.

These values, however, were calculated by choosing arbitrarily the position of the source, or the direction of emission of the sound, so the energy distribution would not be uniform.

Knudsen (1932) carried out experiments designed to show that for rooms of the usual shapes the mean free path of sound is independent of the shape. With the help of a flash-lamp and mirror, he used light rays in place of sound, in scale models of the auditoria considered, and traced the paths followed by successively reflected rays emitted in a limited number of evenly distributed directions, and averaged the distance travelled between reflections, taking the same number of reflections for each emitted ray. His results agree to a good approximation in most cases with Jaeger's formula $4V/S$.

It is the purpose of this paper to show, by direct averaging, that the mean free path of sound in (*a*) *any* rectangular enclosure, (*b*) a spherical enclosure, (*c*) *any* cylindrical enclosure, is $4V/S$.

§2. OUTLINE OF METHOD

When a uniform source of sound has been active in an enclosure for an appreciable time, the sound energy density may be assumed uniform throughout, i.e. the energy is travelling uniformly in all directions from all points in the enclosure (with the exception of certain special cases in which "focusing" occurs). Again, the coefficient of absorption is assumed to be sufficiently low for many reflections to take place, and the sound disturbance between reflections is assumed to travel in straight lines with constant speed until its energy is dissipated.

Now the sound energy may be considered to consist of very small units, or quanta, which move in straight lines with speed c between impacts with the walls, and which do not interfere with each other's motion in any way. (It is not suggested that these quanta are indivisible units.)

Suppose that, in a very small interval of time after the energy density has become uniform in the sense indicated above, $4\pi n$ "quanta" (i.e. n per unit solid angle) leave each of a large number of points which are uniformly distributed throughout the enclosure with volume density ρ.

The method consists in finding the total number of impacts with the interior surfaces of the room made by all these quanta during a finite time T immediately following that short interval, and dividing the total distance $4\pi n\rho cT$, covered by all of the quanta, by this number of impacts. For convenience, the number of impacts, and the distance covered, *per second*, have been used in the following proofs.

§ 3. RECTANGULAR ENCLOSURE

Dimensions of room are a, b, h, with axes of coordinates parallel to edges of room as shown.

Speed of each quantum $= c$.

A quantum projected from any point S in the direction (θ, ϕ) has velocity components $c \sin\theta \cos\phi$, $c \sin\theta \sin\phi$, $c\cos\theta$, in the x, y, z directions.

In the x-direction the distance travelled between impacts with walls is a.

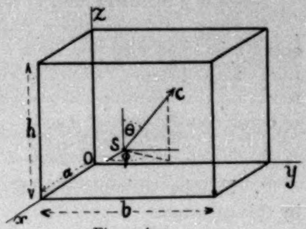


Figure 1.

$\therefore$ Number of impacts per sec. with walls perpendicular to the x-axis is

$$(c/a)\sin\theta\cos\phi.$$

Similarly, numbers of impacts per second with y-walls and z-walls are

$$(c/b)\sin\theta\sin\phi \quad \text{and} \quad (c/h)\cos\theta.$$

$\therefore$ Total number of impacts made per second by one quantum

$$=(c/a)\sin\theta\cos\phi+(c/b)\sin\theta\sin\phi+(c/h)\cos\theta.$$

Now the elementary solid angle with its axis in the (θ, ϕ) direction is

$$\sin\theta \,.\, d\theta \,.\, d\phi,$$

and the number of quanta projected from one point in the (θ, ϕ) direction may therefore be taken as

$$n\sin\theta \,.\, d\theta \,.\, d\phi.$$

$\therefore$ Total number of impacts per second made by quanta from one point is

$$8n \int_0^{\pi/2}\int_0^{\pi/2} \left\{ \frac{c}{a}\sin\theta\cos\phi + \frac{c}{b}\sin\theta\sin\phi + \frac{c}{h}\cos\theta \right\} \sin\theta \,.\, d\theta \,.\, d\phi,$$

and by direct integration this becomes

$$2n\pi\,[c/a+c/b+c/h]$$
$$=2n\pi c \,.\, (bh+ha+ab)/abh.$$

But area of walls $=2(bh+ha+ab)=S$ and volume of room $=abh=V$.

$\therefore$ Number of impacts per second made by quanta from one point

$$=\pi nc \,.\, S/V.$$

The total number of quanta projected from this point is $4\pi n$, and each travels a distance c per second.

∴ Total distance covered per second by these quanta is $4\pi nc$.

∴ Mean distance travelled between impacts

$$=4\pi nc \div \pi nc \,.\, S/V = 4 \,.\, V/S.$$

This is independent of the position of the point of projection, and is, therefore, the same for all such points.

∴ Mean free path $=4V/S$.

§4. SPHERICAL ENCLOSURE

Radius of sphere $=a$.

z-axis is a diameter, and O the centre.

Consider first a quantum projected with speed c in (θ,ϕ) direction from a point S on the z-axis at distance b from the centre. After reflection, this quantum will describe a number of equal chords, each of length

$$2\sqrt{a^2-b^2\sin^2\theta}.$$

∴ Number of impacts made per second by this quantum

$$=c/(2\sqrt{a^2-b^2\sin^2\theta}).$$

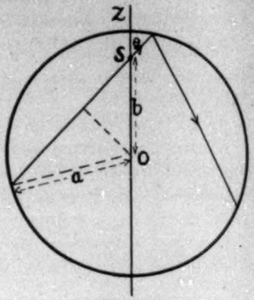


Figure 2.

As before, $n\sin\theta \,.\, d\theta \,.\, d\phi$ quanta are projected into the elementary solid angle whose axis is in the (θ,ϕ) direction.

∴ Total number of impacts made per second by all the $4\pi n$ quanta projected from S

$$=\int_0^\pi\int_0^{2\pi}\frac{nc\sin\theta \,.\, d\theta \,.\, d\phi}{2\sqrt{a^2-b^2\sin^2\theta}}$$

$$=\pi nc\int_0^\pi\frac{\sin\theta \,.\, d\theta}{\sqrt{(a^2-b^2)+b^2\cos^2\theta}}$$

$$=\frac{\pi nc}{b}\left[-\sinh^{-1}\left(\frac{b}{\sqrt{a^2-b^2}}\cos\theta\right)\right]_0^\pi$$

$$=\frac{2\pi nc}{b}\,.\,\sinh^{-1}\frac{b}{\sqrt{a^2-b^2}}$$

$$=\frac{\pi nc}{b}\log_e\frac{a+b}{a-b}.$$

This expression represents the number of impacts made per second by $4\pi n$ quanta.

The total distance covered by them per second is $4\pi nc$.

∴ Mean distance travelled batween impacts by all quanta projected from S is

$$4b\,\log_e\frac{a+b}{a-b}.$$

By substituting a series of values for b, it is seen that the value of this mean free path varies with the point of projection, having values ranging between 0 and $2a$.

If $b \to 0$, the mean free path becomes

$$\underset{b\to 0}{\mathrm{Lt}} \frac{4b}{\frac{2b}{a}+\frac{2}{3}\frac{b^3}{a^3}+\dots},$$

i.e. $$2a \quad \text{or} \quad 6V/S.$$

This agrees with the result obtained by Schuster and Waetzmann for a sphere, and it can be seen that if the source of sound is located at the centre of the sphere all the " rays " will be reflected back to this point, and the uniform distribution of energy will never be attained. This is a special case of " focusing ", and a formula obtained by placing the source at the centre of the sphere cannot be applied to the general case.

If the source is at any point other than the centre, the energy distribution will be uniform when the source has been sounding long enough for a considerable number of reflections to have occurred—the condition assumed throughout—and the points of projection of the " quanta ", as explained above, will then be distributed over the whole volume.

Suppose that there are ρ such points of projection per unit volume. Then the number lying at a distance between b and $b+db$ from O is $\rho \, . \, 4\pi b^2 \, . \, db$.

The number of impacts made per second by quanta projected from one such point has been shown to be

$$\frac{\pi n c}{b} \, . \, \log_e \frac{a+b}{a-b}.$$

$\therefore$ Number of impacts made per second by quanta projected from all points in the enclosure is

$$\int_0^a \frac{\pi n c}{b} \, . \, \log_e \frac{a+b}{a-b} \, . \, \rho \, . \, 4\pi b^2 \, . \, db,$$

which reduces to $4\pi^2 \rho n c a^2 = \pi \rho n c \, . \, S$, where S is the surface area.

But the total number of quanta is $\rho V \, . \, 4\pi n$, so the total distance covered per second is $4\pi\rho n c \, . \, V$.

$\therefore$ Mean free path $= 4\pi\rho n c \, . \, V \div \pi\rho n c \, . \, S = 4V/S$.

§ 5. CYLINDRICAL ENCLOSURE

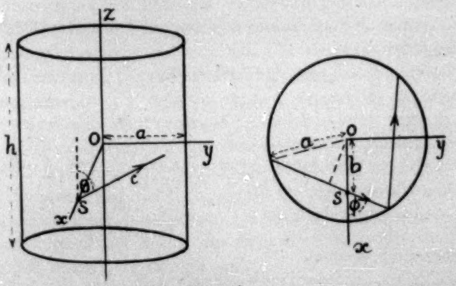


Figure 3.

Cylinder of height h and radius a. Axis of cylinder is z-axis.

Consider a quantum projected with speed c in the (θ, ϕ) direction from the point S, at a distance b from the x-axis. For convenience in reference, let S lie on the x-axis.

Component velocity parallel to the z-axis $= c\cos\theta$.

Distance traversed between impacts in this direction $= h$.

∴ Number of impacts made per second with ends of the cylinder

$$=(c/h)\cos\theta.$$

Component velocity perpendicular to the z-axis, i.e. in x-y plane,

$$=c\sin\theta.$$

Projection of path on this plane consists of a series of equal chords, each of length

$$2\sqrt{a^2-b^2\sin^2\phi}.$$

∴ Number of impacts made per second with curved walls

$$=\frac{c\sin\theta}{2\sqrt{a^2-b^2\sin^2\phi}}.$$

∴ Total number of impacts made per second by one quantum

$$=\frac{c}{h}\cos\theta+\frac{c\sin\theta}{2\sqrt{a^2-b^2\sin^2\phi}}.$$

As before, $n\sin\theta \,.\, d\theta \,.\, d\phi$ quanta are projected into the elementary solid angle with its axis in the (θ, ϕ) direction.

∴ Total number of impacts made per second by quanta projected from S

$$=n\int_0^{\pi}\int_0^{2\pi}\left\{\frac{c\cos\theta}{h}+\frac{c\sin\theta}{2\sqrt{a^2-b^2\sin^2\phi}}\right\}\sin\theta \,.\, d\theta \,.\, d\phi$$

$$=8cn\int_0^{\pi/2}\int_0^{\pi/2}\left\{\frac{\sin 2\theta}{2h}+\frac{1-\cos 2\theta}{4\sqrt{a^2-b^2\sin^2\phi}}\right\}d\theta \,.\, d\phi$$

$$=\frac{2\pi cn}{h}+\pi cn\int_0^{\pi/2}\frac{d\phi}{\sqrt{a^2-b^2\sin^2\phi}}.$$

This expression represents the number of impacts made per second by the $4\pi n$ quanta projected from S.

Suppose, as before, that the points of projection are uniformly distributed with volume density ρ.

∴ The number of such points at distances from the axis of the cylinder between b and $b+db$ is

$$\rho \,.\, 2\pi h \,.\, b \,.\, db.$$

∴ Total number of impacts made per second by quanta projected from all points in enclosure

$$=\int_0^a\left\{\frac{2\pi cn}{h}+\pi cn\int_0^{\pi/2}\frac{d\phi}{\sqrt{a^2-b^2\sin^2\phi}}\right\}\rho \,.\, 2\pi h \,.\, b \,.\, db,$$

$$=2\pi^2\rho cna^2+2\pi^2\rho cnh\int_0^{\pi/2}\int_0^a\frac{d\phi \,.\, b\,db}{\sqrt{a^2-b^2\sin^2\phi}},$$

which on integration becomes

$$\pi\rho cn[2\pi a^2+2\pi ha]=\pi\rho cn \,.\, S,$$

where S is the surface area.

This is the total number of impacts made per second.

But total number of quanta is $\rho V \,.\, 4\pi n$, so that the total distance travelled per second is

$$\rho V \,.\, 4\pi nc.$$

$$\therefore \text{ Mean free path} = 4\pi\rho nc \,.\, V \div \pi\rho nc \,.\, S = 4V/S.$$

REFERENCES

Eyring, 1930. *J. Acoust. Soc. Amer.*, **1**, 217.
Jaeger, 1911. *Wiener Akad. Ber. Math. Naturw. Kl.*, **120**, II *a*.
Knudsen, 1932. *Architectural Acoustics* (London : John Wiley), Chapter V.
Schuster and Waetzmann, 1929. *Ann. Phys., Lpz.*, **1**, 5, 671.

DETERMINATION OF THE CRYSTAL STRUCTURE OF GOLD LEAF BY ELECTRON DIFFRACTION

By T. B. RYMER and C. C. BUTLER*,
The University, Reading
* Now at the University of Manchester

MS. received 10 *April* 1946 ; *in revised form* 15 *January* 1947

ABSTRACT. It is found that the radii of the rings of Debye-Scherrer electron-diffraction photographs obtained from gold leaf are not in exact agreement with the theoretical values. This is ascribed to the crystal lattice being distorted by surface-tension forces.

§ 1. INTRODUCTION

The radii of the Debye-Scherrer rings of electron-diffraction patterns are generally assumed to agree exactly with the predictions of simple theory. So far as we are aware, no attempt has been made to make measurements of higher accuracy than one or two parts in a thousand. Usually, the rings have a radius of the order of one or two cm. and are measured with a travelling microscope having 10-micron divisions ; most observers are satisfied if their readings agree to within a few divisions. There are two main reasons for failure to obtain a higher accuracy. First, plates with a finer grain and microscopes with higher magnification than usual are required. Secondly, if attempts are made to test precision by measuring the radius of a single ring in different azimuths or by comparing the relative radii of different rings, discrepancies are observed which appear to indicate measurement errors of the order of a few parts in a thousand. However, as a result of extensive measurements and stringent statistical tests, we have established that the present practical limit of precision is really of the order of one part in ten thousand, and that the apparent discrepancies alluded to are due to features of the diffraction pattern which are not dealt with by present theories. These peculiarities are of two kinds: the rings are not exactly circular and their radii are not given exactly by Bragg's law.

In a recent paper (Rymer and Butler, 1945 a), we have given a general discussion of the problem of measuring ring radii and have presented some evidence in support of our claim to be able to make measurements of high accuracy. Some of the purely instrumental effects responsible for the abnormal radii and non-circular form of rings are discussed in another paper (Rymer and Butler, 1945 b); the analysis of these features provides an independent check on the precision of measurement.

Our present purpose is to discuss a peculiarity of the diffraction pattern of gold leaf which is revealed by these high-precision measurements: that the relative radii of the rings differ from the expected values by amounts of the order of 0·05 per cent. In view of the crucial importance of an accurate estimate of the precision of measurement, we give a brief description of the experimental technique and we present additional evidence showing the accuracy attained.

§ 2. EXPERIMENTAL

The diffraction camera used was of the type described by Finch, Quarrell and Wilman (1935) with minor modifications. The high-tension supply was a half-wave rectifier set with the output fed through a saturated diode. An additional 500-pF. condenser connected across the camera discharge tube served to reduce voltage fluctuations. Oscillographic examination showed that the ripple voltage was 75 volts r.m.s. at 100 c.p.s. when the discharge tube was taking its normal load of $\frac{1}{2}$ ma. at 50 kv.

Particular care was taken to eliminate stray alternating magnetic fields by the use of compensating coils carrying currents of suitable magnitude and phase. Tests with a search coil and oscillograph showed that the alternating magnetic field at the axis of the camera nowhere exceeded a peak value of $10\,\gamma$. It has been shown (Rymer and Butler, 1945 b) that the slight broadening of the rings introduced by the combined action of alternating fields and high-tension ripple is incapable of introducing errors into the radius measurements of more than $\frac{1}{2}\,\mu$.

Ilford Thin Film Half Tone plates were used with a borax fine-grain developer. The uniformity of this plate and the fineness of the grain combine to make it superior for this purpose to any sofar tried. Plates with coarser grain, such as Ilford Ordinary or Special Rapid, are quite unsuitable for precision work. Tests show that with a constant intensity of electrons of about 50 kv. energy, the density produced on a Thin Film plate is accurately proportional to the exposure time up to densities of at least unity; since the reciprocity law holds for electrons (Becker and Kipphan, 1931), the density is proportional to the intensity. This is a matter of importance, since otherwise systematic errors can occur in the measurement of the ring radius (Rymer and Butler, 1945 a).

The width of the beam where it strikes the photographic plate has been measured (Rymer and Butler, 1945 b) and found to be $40\,\mu$. The width of the diffraction rings produced by gold leaf is of the order of $200\,\mu$, and is therefore due almost entirely to broadening by the crystal; there is therefore little to be gained by further attempts to sharpen the electron beam. That the broadening is due to the finite size of the crystals rather than to a variable lattice constant is suggested by the data of table 1, which show that the ring width is independent of the radius.

Table 1. Plate C/279. Gold-leaf specimen. Widths of diffraction rings

Indices	111	200	220	311
Width (microns)	140	116	110	130

Measurements of the ring radii were made with an instrument reading to one micron (Rymer and Butler, 1944) which can be used as a travelling microscope or as a non-recording microphotometer and is fitted with a divided head which permits azimuths of radii measurements to be determined to 0°·5. Corrections have to be applied to the measured radii to allow for the following effects:

(*a*) Errors of microphotometer screw.

(*b*) Background density due to incoherent scattering of electrons.

(*c*) Curvature effect: the intensity is enhanced on the inner side of a diffraction ring owing to its shorter perimeter.

(*d*) Finite length of microphotometer slit.

The method of determining these corrections has previously been described (Rymer and Butler, 1945 a).

§ 3. PRECISION OF MEASUREMENT

Before presenting the results of measurements on gold specimens, it is desirable to discuss the evidence of the accuracy of the technique provided by substances which do not exhibit any peculiarities. Our belief in the reliability of our measurements is based on the agreement between the values of the standard deviation as computed by the following completely independent methods:

(*a*) Analysis of the scatter of repetition readings.

(*b*) Analysis of the scatter of measurements of the radius of a single ring in different azimuths when allowance is made for instrumental effects.

(*c*) Comparison of the variation of the ellipticity of the diffraction rings due to stray magnetic fields with theoretical predictions.

(*d*) Comparison of the mean radii of diffraction rings with theoretical values.

This is illustrated by measurements made on Plate No. D/137 of a sodium chloride specimen (figure 4).

(*a*) The plate was measured with the microscope by making three settings on each ring and repeating the process in each of eighteen equally spaced azimuths. From the scatter of each group of three settings from their mean, the standard deviation of a single setting on the (111), (200), (220) and (420) rings is found to be 2·6, 2·2, 2·2 and 2·9 μ respectively. The average standard deviation of a single setting is thus

$$\text{S.D.} = 2{\cdot}49 \pm 0{\cdot}15\,\mu.^{*} \qquad \ldots\ldots(1)$$

(*b*) The radius of the (200) ring was measured in azimuths 0, 20, 40, ... 340 degrees using the microscope, the mean of three readings being taken in each case. Table 2 shows the results, the radii being given in microns.

Table 2. Radius of (200) ring = 8790 + following

Azimuth	0	20	40	60	80	100	120	140	160
Radius	33	26	21	18	9	5	10	22	25
Azimuth	180	200	220	240	260	280	300	320	340
Radius	37	41	53	60	65	68	51	50	39

---

[*] Throughout this paper, ± indicates *standard deviation*.

Fitting a Fourier series to these results gives *

$$\text{Radius} = 8825{\cdot}20 - 26{\cdot}40 \sin(\theta + 1{\cdot}78)^\circ + 2{\cdot}49 \sin(2\theta + 327{\cdot}79)^\circ. \qquad \ldots\ldots(2)$$

The first harmonic is trivial, being due to measurements being made from a point which is not quite the centre of the pattern. The second harmonic is due to the presence of stray magnetic fields (Rymer and Butler, 1945 b). From the differences between the radii given by equation (2) and the numbers in table 2, we obtain the standard deviation of any radius measurement as $3{\cdot}75 \pm 0{\cdot}73\,\mu$. Since each radius measurement is the mean of three readings, the standard deviation of a single reading must be $\sqrt{3}$ times this:

$$\text{S.D.} = 6{\cdot}49 \pm 1{\cdot}27\,\mu. \qquad \ldots\ldots(3)$$

(*c*) Theoretically, the amplitude of the second harmonic in the expression for the radius should be proportional to the mean radius, while the phase should be the same for all rings (Rymer and Butler, 1945 b). Table 3 gives the amplitude and phase of the harmonic for four rings.

Table 3

Indices	Amplitude (microns)	Phase (°)	Mean radius (microns)	$\frac{\text{Amplitude}}{\text{Radius}}$
111	3·2	244·1	7645·6	$4{\cdot}19 \times 10^{-4}$
200	2·5	327·8	8827·1	2·83
220	7·0	290·0	12477·0	5·61
420	8·7	280·3	19731·3	4·41

The larger harmonics naturally give more accurate values for the amplitude/radius. We therefore weight the results in proportion to the radius, obtaining for the mean amplitude/radius $4{\cdot}40 \times 10^{-4}$. From the differences between this and the numbers in the fifth column of table 3 it can be deduced that the standard deviation of a single harmonic is $1{\cdot}18 \pm 0{\cdot}48\,\mu$. Now it can be shown that the standard deviation of the amplitude of a harmonic computed from N radii measurements is $\sqrt{(2/N)}$ times the standard deviation of a single radius measurement. Since there are 18 radii measurements, and each is the mean of three readings, we obtain for the standard deviation of a single reading

$$\text{S.D.} = 6{\cdot}15 \pm 2{\cdot}51\,\mu. \qquad \ldots\ldots(4)$$

An estimate of the precision can also be made from the phase angles of the second harmonic. The weighted mean phase is 285°·72. Now the standard deviation of the phase (measured in radians) is equal to the standard deviation of the amplitude of the harmonic divided by the amplitude. Using this relation, it can be calculated from the data of table 3 that the standard deviation of the amplitude of a single harmonic is $2{\cdot}23 \pm 0{\cdot}91\,\mu$, corresponding to a standard deviation of a single reading of

$$\text{S.D.} = 11{\cdot}60 \pm 4{\cdot}73\,\mu. \qquad \ldots\ldots(5)$$

---

* It must be emphasized that the fact that the coefficients in this series are given to the second decimal place does *not* imply that their value is necessarily known even to the first decimal place. The coefficients have been calculated to the second decimal place merely in order to avoid any possibility of "rounding-off" errors in the determination of the standard deviation. Similar remarks apply to *all figures quoted in this paper*. The *only* valid estimate of accuracy is based on a calculation of the standard deviation.

(*d*) If d is the interplanar spacing, R the radius of the diffraction ring, L the distance from specimen to photographic plate and λ the wave-length of the electrons, then theoretically

$$Rd = \lambda L + \tfrac{3}{8}\frac{\lambda L}{L^2}R^2. \qquad \ldots\ldots(6)$$

For several reasons this relation does not hold exactly, but the values of λL calculated from this equation vary systematically with the ring radius according to the relation (Rymer and Butler, 1945 a and b)

$$\lambda L = (\lambda L)_0 + B/R^2, \qquad \ldots\ldots(7)$$

where B is a constant. In our first paper (1945 a) we have shown that the experimental results for the plate under discussion fit these equations, the standard deviation of a single λL determination being $2{\cdot}2 \times 10^{-12}\,\text{cm}^2$, while $(\lambda L)_0 = 2{\cdot}48093 \times 10^{-8}\,\text{cm}$. Since the average radius R is 1·2170 cm., this gives for the standard deviation of a mean radius 1·09 μ. A rather more satisfactory procedure is to weight the (λL) determinations in proportion to the radius. When this is done, and the constants of equation (7) are determined by the method of least squares, it is found that the standard deviation of the mean radius of a ring is $0{\cdot}87 \pm 0{\cdot}43\,\mu$. This is made up of two parts: random errors in the individual radii measurements and errors in the determination of the systematic corrections. If the latter are neglected, we obtain for the standard deviation of a single reading

$$\text{S.D.} = 6{\cdot}39 \pm 3{\cdot}19\,\mu. \qquad \ldots\ldots(8)$$

This is almost certainly too high as the error in the systematic corrections is unlikely to be entirely negligible. A reasonable estimate (Rymer and Butler, 1945 a) of this latter error is 0·77 μ. Using this value, the standard deviation of a single reading becomes

$$\text{S.D.} = 2{\cdot}99 \pm 1{\cdot}49\,\mu. \qquad \ldots\ldots(9)$$

The higher value of the standard deviation derived from the study of the amplitude and phase of the second harmonic (equations (4) and (5)) may be due to the stray magnetic fields not being *linear* functions of position as is assumed in the theory which predicts that the amplitude should be proportional to the radius and that the phase should be the same for all rings. Unpublished results from a large number of photographs of many substances indicate that in general the diffraction rings are not circular and that the expression for the radius given by equation (2) should be extended by the addition of higher harmonics whose origin will be discussed in a later paper. Owing to the presence of these harmonics, the estimate of the standard deviation given in equation (3) is too high. Comparing the results for the standard deviation of a single reading given by equations (1), (3), (4), (5) and (9), we see that this quantity is certainly less than 6 μ and that a very fair estimate is 3 μ.

Naturally, few plates have beeen studied so exhaustively as the one discussed. Less complete measurements have been made on many plates for the study of special features, and these all confirm the estimate given of the precision of measurement for rings of this width. Since the width of the rings and general quality of the plate under discussion is very little different from that of the gold-leaf diffraction patterns to be discussed (compare figure 4 with figures 5 and 6), we may take 3 μ as a reasonable estimate of the standard deviation of a single reading for the latter plates.

§4. GOLD-LEAF DIFFRACTION PATTERNS

When an attempt is made to fit equations (6) and (7) to measurements of gold-leaf diffraction patterns, discrepancies are immediately observed. The nature of these can be illustrated by the results for a typical plate (figure 5) of a diffraction pattern of a specimen of gold leaf which had been thinned in potassium cyanide solution. Table 4 gives the mean radius R of the first four diffraction rings and the values of λL calculated from them by means of equation (6), the lattice constant of gold being taken as 4·0700 A. and the camera length L as 47 cm. Figure 1 is a graph of λL against $1/R^2$. The dotted line has been fitted by the method of least squares, the different points being given weights proportional to R, and has the equation (R in cm.)

$$\lambda L = (22836{\cdot}47 + 13{\cdot}77/R^2) \times 10^{-12}\,\text{cm}^2 \qquad \ldots\ldots(10)$$

The values of λL calculated from this are listed in table 4 under $(\lambda L)_{\text{cal.}}$ (eqn. (10)),

Table 4. Plate No. D/268

Indices	R (microns)	λL 10^{-12} cm²	$(\lambda L)_{\text{calc.}}$ (eqn. (10))	$(\lambda L)_{\text{calc.}}$ (eqn. (11))
111	9724·8	22847·8	22851·03	22847·67
200	11232·4	53·0	47·38	44·94
220	15879·0	39·5	41·93	40·84
311	18623·9	40·8	40·44	39·62

and from the differences between these and the observed λL it can be calculated that the standard deviation of each radius determination is $2{\cdot}50 \pm 1{\cdot}25\,\mu$. Of this amount, $0{\cdot}77\,\mu$ arises from the error in the systematic corrections (Rymer and Butler, 1945 a). Also, a radius determination is the mean of 54 readings, for the plate was measured along 18 equally spaced azimuths and in each case three settings were made on the rings. Hence the standard deviation of a single setting must be $\sqrt{54}\,\sqrt{2{\cdot}50^2 - 0{\cdot}77^2} = 17{\cdot}5 \pm 8{\cdot}8\,\mu$. This is distinctly greater than the expected value of $3\,\mu$.

Examination of figure 1 suggests that the points corresponding to the (111), (220) and (311) rings lie much more nearly on a straight line than do all four points. The full line has been fitted to these points by the method of least squares and has the equation

$$\lambda L = (22836{\cdot}73 + 10{\cdot}35/R^2) \times 10^{-12}\,\text{cm}^2 \qquad \ldots\ldots(11)$$

The values of λL calculated from this are listed in table 4 under $(\lambda L)_{\text{calc.}}$ (eqn. 11), and from the differences between them and the observed λL it can be calculated that the standard deviation of a radius determination (allowing for errors of the systematic corrections) is $1{\cdot}03\,\mu$, corresponding to a standard deviation of a single setting of $7{\cdot}5 \pm 5{\cdot}3\,\mu$. This is much closer to the expected value. On the other hand, equation (11) implies that the observed value of λL for the (200) diffraction is $(8{\cdot}07 \pm 2{\cdot}76) \times 10^{-12}$ cm². larger than would be expected.

We have so far measured over a dozen diffraction patterns of gold leaf prepared under various conditions, and in *every* case the value of λL calculated from the (200) diffraction ring is higher than would be expected from the results for the remaining rings. In table 5 the magnitude of this anomaly is given in the third column. The form of the variation of the anomaly with the wave-length of the electrons is

Table 5. (200) diffraction anomaly

Plate No.	$(\lambda L)_0$ $(10^{-12}$ cm$^2)$	Anomaly $(10^{-12}$ cm$^2)$	$\frac{\text{Anomaly}}{(\lambda L)_0}$ $(\times 10^{-4})$	Remarks
D/257	22510·32	8·73	3·88	1938 gold. Amalgam rings
D/268	22836·73	8·07	3·53	1945 gold.
D/258	23349·32	7·21	3·09	1938 gold. Amalgam rings
D/117	23738·42	9·04	3·81	1938 gold. Annealed
C/251	25454·40	12·25	4·81	1938 gold.
C/276	25742·35	11·26	4·37	1938 gold.
D/269	29847·74	10·98	3·68	1938 gold.
D/261	30282·10	12·72	4·20	1945 gold.

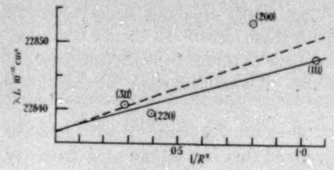


Figure 1.

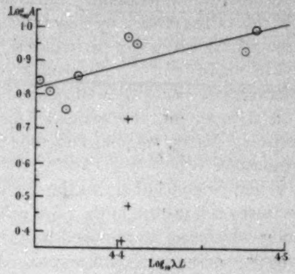


Figure 2.

shown in figure 2, which is a graph of the logarithm of the anomaly against the logarithm of the corresponding $(\lambda L)_0$. The straight line has been fitted by the method of least squares and has the equation

$$\log_{10} A = -5{\cdot}174 + (1{\cdot}40 \pm 0{\cdot}42) \log_{10} (\lambda L)_0. \qquad \ldots\ldots(12)$$

Within the limits of experimental error, the anomaly is *proportional* to $(\lambda L)_0$. Values for the ratio of the two are listed in the fourth column of table 5; their mean value is

$$\text{Anomaly}/(\lambda L)_0 = (3{\cdot}92 \pm 0{\cdot}20) \times 10^{-4}. \qquad \ldots\ldots(13)$$

The gold specimens were prepared from two samples of gold leaf purchased in 1938 and 1945 respectively. Comparison of table 5 with figure 2 shows that there is no perceptible difference between the results for the two samples. An analysis of the latter sample showed it to contain 0·09% copper and 0·02% silver. As it is not possible to obtain gold leaf of greater purity than this, the effect of impurity was examined by obtaining diffraction patterns from a less pure gold leaf containing 2·55% silver and 1·35% copper. The results for plates from this are given in table 6 and are plotted in figure 2 as crosses. It is clear that the addition of impurity, far from being the cause of the anomaly, actually tends to reduce it.

Table 6. Results for impure gold leaf

Plate No.	$(\lambda L)_0$ $(10^{-12}$ cm$^2)$	Anomaly $(10^{-12}$ cm$^2)$
D/293	25534·2	3·7
D/294	25290·8	3·0
D/305	25460·4	7·0

We have attempted to investigate this point further by obtaining diffraction patterns from gold films prepared by electro-deposition by the method of Finch and Sun (1936); such films might be expected to be more free from impurities than the best commercial gold leaf. There was an indication that the (200) anomaly was somewhat larger for such specimens, but it was impossible to be certain as the crucial (200) ring was always weak and the determination of the correction for background density consequently liable to large error.

The majority of the results of table 5 and figure 2 were obtained from gold leaf which had not been annealed. Plate D/117, however, was obtained from a specimen which had been annealed at 340° c. for 22 hours after thinning in potassium cyanide solution. Since the result for this plate is not sensibly different from that of the others, we may conclude that the anomaly is not due to any strain in the crystal which can be removed by annealing at this temperature.

The majority of the plates showed no trace of any amalgam rings (figure 5). However, two plates showed marked amalgam rings (figure 6). Since the values of the anomaly for these plates are not perceptibly different from its values for the remaining plates, we may conclude that traces of amalgamation can hardly be responsible.

It may be concluded that the (200) diffraction anomaly observed with gold-leaf specimens is a feature of the pure gold lattice, and that it does not arise from any strains which can be removed by annealing. It would have been of interest to examine specimens of gold evaporated on to cellulose; unfortunately, the diffraction rings obtained from such specimens are so broad that it is impossible to make measurements of sufficiently high precision.

As an additional test, measurements have been made on patterns from a specimen the plane of which was not perpendicular to the electron beam. Table 7 shows that such a tilted specimen yields a value for the ratio of the anomaly to $(\lambda L)_0$ which is not sensibly different from the average value (3·92) for the other plates.

Table 7. Effect of tilting specimen

Plate No.	Angle of tilt (°)	$(\lambda L)_0$ (10^{-12} cm^2)	Anomaly (10^{-12} cm^2)	$\frac{\text{Anomaly}}{(\lambda L)_0}$
D/280	65	23126·0	7·75	$3{\cdot}35\times10^{-4}$

In this section, we have regarded our results as indicating that the value of (λL) calculated from the (200) diffraction ring is higher than would be expected. Consideration of figure 1 shows that this is not the only possible explanation. The small difference in the abscissae of the points representing the (220) and (311) rings makes it equally possible to regard these two and the (200) point as lying on a straight line, in which case the value of λL for the (111) ring is abnormally low. According to the theory presented in the next section, there is no essential difference between these two interpretations.

§ 5. INTERPRETATION OF ANOMALY

The anomalies discussed in the preceding section might be attributed to one or more of the following causes: (*a*) refraction of the electrons at the surface of the gold crystallites, (*b*) variation of the angle of diffraction from the Bragg value in accordance with the dynamical theory, (*c*) deformation of the crystals from exact cubic form.

(*a*) The effect of refraction can be shown to be two-fold: the rings are broadened, or even split into doublets (Sturkey and Frevel, 1945), and the peak of the broadened ring is displaced. Such a displacement can be shown to result in an anomaly in the value of λL proportional to the *cube* of the electron wave-length. Now the experimental results as represented by equation (12) and figure 2 are consistent with a first-power law but definitely rule out the possibility of a cube law. The origin of the anomaly cannot therefore be sought in refraction effects.

(*b*) The fact that the intensities of the (111) and (200) rings of gold are considerably less than is required by the kinematic theory (Tol and Ornstein, 1940) suggests that dynamical effects are very marked. Nevertheless, the following considerations indicate that such effects are probably not the main cause of the observed anomalies. In the first place, Thomson and Blackman (1939) have shown that for transmission through a parallel-sided slab of crystal ("Laue case") the dynamical theory predicts a change in the angle of diffraction by an amount $\zeta\lambda \tan\psi/2\pi$, where ζ is the *resonance error* of Bethe's theory, λ the wave-length of the electrons and ψ the angle between the reflecting plane and the surface of the crystal. According to the dynamical theory, ζ ranges between approximately

$$\pm \frac{v\lambda}{2\pi},$$

where v is the Fourier coefficient corresponding to the diffraction, so that the angle of diffraction can deviate from the normal by amounts up to

$$\pm \frac{v\lambda^2 \tan\psi}{4\pi^2}.$$

The elementary dynamical theory therefore predicts no change in the mean radius of the diffraction ring but only a *symmetrical broadening*, which should be of the order of 200 μ for a camera length of 50 cm. Nevertheless, since the observed anomalies amount to only some 4% of the ring width, it is not inconceivable that a refinement of the present dynamical theory might account for them as a higher-order effect. Since the elementary dynamical theory leads to symmetrical deviations in the angle of diffraction proportional to λ^2, it would be expected that any such higher-order effect would be proportional to at least the second power of λ, while the experimental results (equation (12)) show that even a second-power dependence on λ is less likely than a first power, and any higher power than the second is definitely ruled out. We conclude that the possibility of the anomalies being due to dynamical effects cannot be entirely ignored, but that it is not very likely, and in any case cannot be further discussed, in the present state of the theory.

(*c*) If the anomalies are due to a departure of the crystals from cubic symmetry, they would be proportional to the *first* power of the electron wave-length, in agreement with equation (12). However, the assumption of a non-cubic lattice is not, of itself, sufficient to explain the results. If, for example, it be supposed that the crystal lattice is slightly deformed from cubic to tetragonal form without change of volume of the unit cell, then one of the (200) spacings is increased (decreased) while the other two are decreased (increased). The diffraction pattern from a random arrangement of such crystals would give a broadened

(200) ring but with its peak undisplaced. Similar reasoning can be applied to other planes of the crystal and to other assumed deviations from cubic form. The only way in which a pseudo-cubic crystal lattice could give the observed results is if there is some orientation of the crystallites so that—to take the example just given—the (200) planes of increased spacing are never approximately parallel to the electron beam. We must, however, reject this explanation for two reasons: (i) the degree of orientation as judged by the intensity of the rings and the amount of "arcing" when the specimen is tilted is small and varies considerably in magnitude from specimen to specimen, whereas the magnitude of the anomaly is consistent from one specimen to another; (ii) inclining the specimen to the electron beam makes no appreciable change in the value of the anomaly (see table 7). It therefore appears impossible to explain the results by assuming an ordinary unstressed crystal lattice with flat crystal planes; we are driven to postulate a *stressed* lattice in which the planes have been warped. Such a warping will generally occur when a gold crystal is stressed owing to the fact that it is elastically anisotropic.

While the evidence thus points to stresses in the gold crystals as the cause of the observed anomalies, it does not suffice to determine unambiguously the nature and origin of these stresses; it is found that stress systems of a rather general type are capable of explaining quantitatively the observed results.

Consider the effect of a *simple tension* p in the specimen in a direction making an angle ϕ with the electron beam. Referred to the crystal axes of a certain crystallite, let $\gamma_1, \gamma_2, \gamma_3$ be the direction-cosines of the electron beam and C_1, C_2, C_3 those of the normal to a set of reflecting planes. Then the fractional extension of the crystal along C_1, C_2, C_3 is:

$$\delta = p\left[\frac{-c_{12}}{(c_{11}+2c_{12})(c_{11}-c_{12})} + \frac{C_1^2\gamma_1^2+C_2^2\gamma_2^2+C_3^2\gamma_3^2}{c_{11}-c_{12}} + \frac{C_1C_2\gamma_1\gamma_2+C_2C_3\gamma_2\gamma_3+C_3C_1\gamma_3\gamma_1}{c_{44}}\right], \qquad \ldots\ldots(14)$$

where c_{11}, c_{12}, c_{44} are the usual elastic constants.*

The fractional extension along the [111] direction of the crystals producing the (111) diffraction ring may be found as follows. We have

$$C_1 = C_2 = C_3 = 1/\sqrt{3}.$$

Hence

$$C_1^2\gamma_1^2 + C_2^2\gamma_2^2 + C_3^2\gamma_3^2 = \tfrac{1}{3}. \qquad \ldots\ldots(15)$$

Now let the normal to the (111) planes of a diffracting crystal (this normal must necessarily be very nearly perpendicular to the electron beam) make an angle θ with the projection of the tension p on to a plane perpendicular to the electron beam; the corresponding portion of the (111) diffraction ring is in azimuth θ with respect to the projection of p on the photographic plate. Then the angle between p and the [111] direction is $\cos^{-1}(\sin\phi\cos\theta)$. Hence

$$C_1\gamma_1 + C_2\gamma_2 + C_3\gamma_3 = \sin\phi\cos\theta.$$

whence from (15)

$$C_1C_2\gamma_1\gamma_2 + C_2C_3\gamma_2\gamma_3 + C_3C_1\gamma_3\gamma_1 = \tfrac{1}{4}\sin^2\phi - \tfrac{1}{6} + \tfrac{1}{4}\sin^2\phi\cos 2\theta. \qquad \ldots\ldots(16)$$

---

* *Handbuch der Physik*, 1928, **6**, 418 (Berlin: Springer).

Substituting from (15) and (16) in (14),

$$\delta = p\left[A + \tfrac{1}{3}B + \frac{1}{4c_{44}}C\right], \qquad \ldots\ldots(17.1)$$

where

$$A = \frac{-c_{12}}{(c_{11}+2c_{12})(c_{11}-c_{12})} + \frac{\sin^2\phi}{2(c_{11}-c_{12})},$$

$$B = (1-\tfrac{3}{2}\sin^2\phi)\left(\frac{1}{c_{11}-c_{12}} - \frac{1}{2c_{44}}\right),$$

$$C = \sin^2\phi\cos 2\theta.$$

By an extension of this method, it can be shown that the corresponding expressions for the other rings are all of the form

$$\delta = p[A + \alpha B + \beta C], \qquad \ldots\ldots(17.2)$$

where A, B and C are the same for all rings. The coefficients α and β for the first four rings are given in table 8.

Table 8

Ring indices	$1/(h^2+k^2+l^2)$	α	β
111	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{4c_{44}}$
200	$\frac{1}{4}$	0	$\frac{1}{2(c_{11}-c_{12})}$
220	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8(c_{11}-c_{12})}+\frac{3}{16c_{44}}$
311	$\frac{1}{11}$	$\frac{19}{121}$	$\frac{32}{121(c_{11}-c_{12})}+\frac{57}{484c_{44}}$

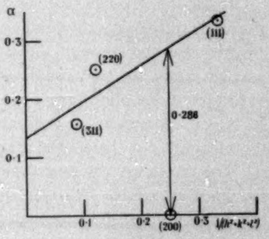


Figure 3.

Now the first term of (17) represents a uniform extension of the crystals in all directions and corresponds merely to a change of scale of the diffraction pattern; it is therefore irrelevant for the present results. The last term gives rise to an ellipticity of the rings, and will be considered later. The second term vanishes in the case of elastically isotropic crystals for which the Cauchy relation $c_{11}-c_{12}=2c_{44}$ is valid, while for anisotropic crystals it gives rise to changes in the relative radii of the different diffraction rings.

If a certain interplanar spacing is *increased* by a small fraction δ, the radius of the corresponding diffraction ring is *decreased* by δ and the value of (λL) computed from it by means of equation (6) using the normal interplanar spacing is likewise decreased by a fraction δ. Hence the ordinates of a $(\lambda L, 1/R^2)$ graph such as figure 1 are displaced by amounts proportional to $-\alpha$. The abscissae of figure 1 are the quantities $1/R^2$, which are very approximately proportional to d^2, i.e. to the numbers in the second column of table 8. The anomaly in the (200) reflection may therefore be evaluated by plotting α against $1/(h^2+k^2+l^2)$ (figure 3), fitting the best straight line to the points corresponding to the (111), (200) and (311) reflections and observing the distance of the (200) point from this line. It is found that this displacement is 0·288. Comparing this with the observed value of (anomaly)/λL (equation 13), we get

$$0{\cdot}288pB = 3{\cdot}92 \times 10^{-4}$$

whence

$$p(1 - \tfrac{3}{2}\sin^2\phi)\left(\frac{1}{c_{11}-c_{12}} - \frac{1}{2c_{44}}\right) = 1{\cdot}36 \times 10^{-3}.$$

Inserting the known values of the elastic constants (Goens and Weerts, 1936),

$$c_{11} = 18{\cdot}6 \times 10^{11} \text{ dynes/cm}^2$$
$$c_{12} = 15{\cdot}7,$$
$$c_{44} = 4{\cdot}2,$$

we obtain

$$p(1 - \tfrac{3}{2}\sin^2\phi) = 6{\cdot}01 \times 10^{8} \text{ dynes/cm}^2$$

The anomaly can therefore be explained if we suppose that there is a tension of $6{\cdot}01 \times 10^8$ dynes/cm^2 parallel to the electron beam ($\phi = 0$); such a tension causes a displacement of the (200) point from the line fitting the other three points, while the displacements of the latter are negligible.

It should be noted that we cannot compare the *slope* of the line of figure 3 with the slope of the experimental curve of figure 1, for it is known (Rymer and Butler, 1945 b) that the slope of the latter is in part due to a charging up of the photographic plate by the undiffracted beam. The magnitude of this charging-up varies from one photograph to another and cannot therefore be easily allowed for. Its effect is to add to the ordinates of the points quantities proportional to $1/R^2$; it therefore cannot change the magnitude of the anomaly.

The stress system postulated above is not a unique solution to the problem, for since a uniform hydrostatic compression reduces *all* interplanar spacings by the same fraction,* a stress system consisting of a simple tension together with an arbitrary hydrostatic pressure will fit the experimental results equally well. In particular, a stress system consisting of a tension p and a hydrostatic compression of the same magnitude is a possible solution. This reduces to a two-dimensional compression in a plane perpendicular to the electron beam. Such a stress system could arise from surface-tension forces if the specimen is in the form of laminae set normal to the beam (i.e. in the plane of the specimen); the interior of a lamina of thickness t of material with a surface tension S will experience a compression in its plane of magnitude $2S/t$. There is no information as to the surface tension of solid gold, and the published values for the molten metal range from 500 to 1000 dynes/cm. If we take the former value, we find $t = 1{\cdot}7 \times 10^{-6}$ cm. as the thickness of a lamina. This is of the order of magnitude of the thickness of transmission specimens.

Equation (17) predicts that when the laminae are not perpendicular to the electron beam ($\phi \neq 0$), two effects should be observed: (i) the magnitude of the anomaly, which is proportional to B, should be reduced by a factor $1 - \frac{3}{2}\sin^2\phi$; (ii) the term βC no longer vanishes, indicating that the rings become elliptical. However, the results of table 7 show that inclining the specimen to the beam does not change the value of the anomaly. Investigation of the ellipticity of the rings is hampered by the unavoidable presence of stray magnetic fields, to which reference has been made in § 3, but the results (which need not be given in detail) show that the ellipticity attributable to strain is not statistically significant, and in any event is smaller by a factor of 10 than is predicted by (ii). These results can be brought into harmony with the stress theory if it be supposed that in the region of the specimen irradiated by the beam there is a large number of domains with a different direction of the tension in each. The ellipticity of the rings would be averaged out, while the (200) anomaly would be that corresponding to the average value of $\sin^2\phi$. Such an effect could be produced by surface-tension forces if the specimen consisted of a number of laminae in random orientation and having a thickness of the order of 4×10^{-7} cm. (assuming a surface tension of 500 dynes/cm.). A specimen in the form of filaments of radius of this order and in random orientation would equally give rise to the observed effects.

---

* *Handbuch der Physik*, 1928, **6**, 418 (Berlin: Springer).

It is apparent from this discussion that the results are consistent with a wide variety of stress systems, and it might therefore be expected that the stresses in the neighbourhood of lattice imperfections such as dislocations or twinning planes would give rise to the observed effects. There are, however, two difficulties in attributing the results to this cause. First, in the neighbourhood of a lattice imperfection there are two regions of equal and opposite stress, and the diffraction rings from these would be displaced by equal and opposite amounts: the resultant ring would be slightly broadened but would not be displaced. Secondly, it would be expected that the neighbourhood of a dislocation would be characterized by a *strain* which would not be sensitive to small traces of impurity, for the essential feature of a dislocation is that a group of atoms is displaced through a distance determined by the lattice constant. The *stress* associated with a dislocation will of course be sensitive to traces of impurity. Now it is found experimentally (compare tables 5 and 6) that the addition of 2·6% of silver greatly reduces the (200) anomaly. This implies a *stress* which is independent of traces of impurity and a *strain* which is diminished when the lattice is hardened by the presence of foreign atoms. This is consistent with a surface tension rather than a dislocation origin of the stress system. Another fact pointing in the same direction is that the magnitude of the (200) anomaly is unchanged by annealing, though this is not conclusive as we were unable to use temperatures above 340° c. without risk of damage to our specimens.

A surface-tension origin of the stress system means that the magnitude of the (200) anomaly depends on the thickness of the diffracting particles, whereas the results of table 5 show that it is very consistent from one specimen to another. It is to be expected that particles of a wide range of thickness will be present. Particles thicker than a certain amount will contribute little to the pattern owing to excessive absorption of the beam. Considerations of mechanical strength will set a lower limit to the size of the particles, and also the smallest particles will have insufficient scattering power to produce a good pattern. The bulk of the diffraction pattern will therefore come from particles of a rather restricted range of thickness, and this probably accounts for the consistency of the observed anomaly from one specimen to another.

§6. ACKNOWLEDGMENTS

The authors gratefully acknowledge the encouragement they have received from Professor J. A. Crowther and the facilities placed at their disposal. The senior author is also indebted to the Research Board of the University of Reading for a grant for the purchase of the electron-diffraction camera.

REFERENCES

Becker, A. and Kipphan, E., 1931. *Ann. Phys., Lpz.*, **10**, 15.
Finch, G. I., Quarrell, A. G. and Wilman, H., 1935. *Trans. Faraday Soc.*, **31**, 1051.
Finch, G. I. and Sun, C. H., 1936. *Trans. Faraday Soc.*, **32**, 852.
Goens, E. and Weerts, J., 1936. *Phys. Z.*, **37**, 321.
Rymer, T. B. and Butler, C. C., 1944. *Phil. Mag.*, **35**, 202; 1945 a. *Ibid.*, **36**, 515; 1945 b. *Ibid.*, **36**, 821.
Sturkey, L. and Frevel, L. K., 1945. *Phys. Rev.*, **68**, 56.
Thomson, G. P. and Blackman, M., 1939. *Proc. Phys. Soc.*, **51**, 425.
Tol, T. and Ornstein, L. S., 1940. *Physica*, **7**, 685.

A COLORIMETER WITH SIX MATCHING STIMULI

By R. DONALDSON,
National Physical Laboratory, Teddington

MS. received 19 *November* 1946

ABSTRACT. The instrument is a modification of the ordinary trichromatic colorimeter. The three matching stimuli of the ordinary instrument, the red, green and blue, have been increased to six by the addition of an orange, yellow-green and blue-green. The spectral energy distribution of the colour being measured is first approximately matched by means of a mixture of all six colours before the final colour match is made by varying three of the colours only. This eliminates to a large extent the personal error of the observer, and allows a large field to be used with a resultant gain in sensitivity.

§1. INTRODUCTION

In the measurement of colour there is naturally a tendency to pass from visual to photoelectric methods. This change-over, however, is not taking place as smoothly as might be expected. The difficulty is that there has not yet appeared a simple photoelectric design which will permit the construction of a cheap reliable instrument. There have been two main lines of development—the spectrophotometer, and the photoelectric colorimeter employing a spectrum template. The former would seem, for the present, to have reached a culmination in the Hardy automatic instrument, and the latter, although it has not received serious attention for such a long time as the spectrophotometer, has already produced two versions of promise (Knipe and Reid, 1943; Winch, 1946).

When considering the obvious advantages of photoelectric methods, the considerable increase of complexity in the apparatus must not be overlooked. A photoelectric instrument, which measures colour accurately and quickly, is

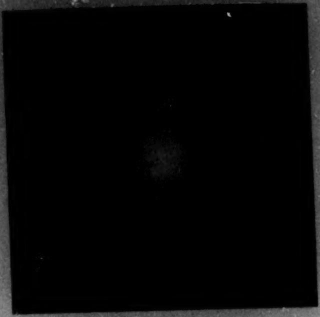


Figure 4. Plate D/137.
Sodium chloride specimen.

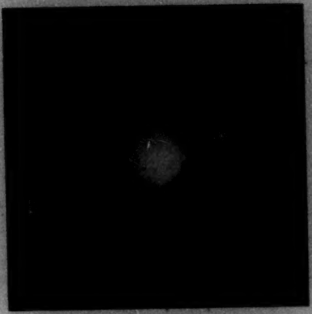


Figure 5. Plate D/268.
Gold-leaf specimen.

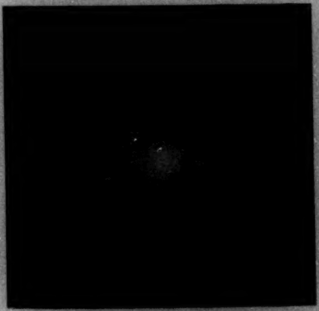


Figure 6. Plate D/258.
Gold-leaf specimen showing amalgam rings.

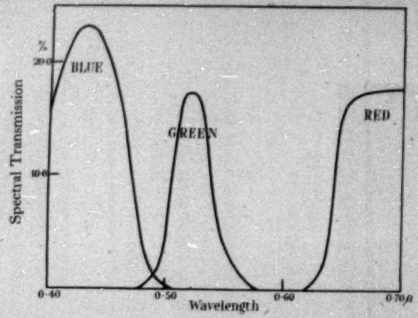


Figure 1. Mixture colours of the ordinary trichromatic colorimeter.

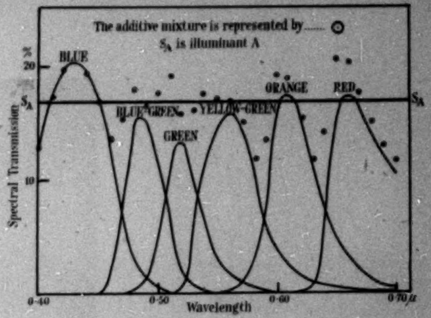
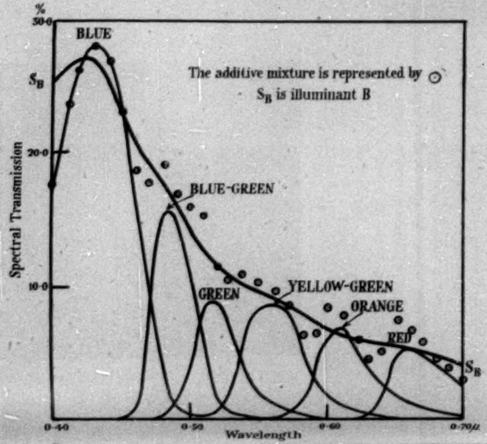
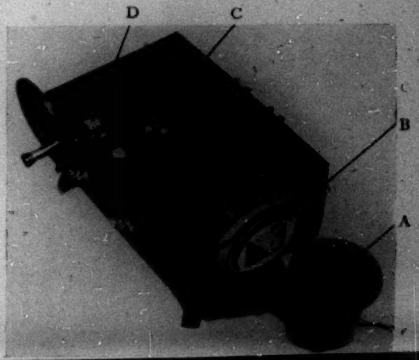


Figure 2. Energy match with standard illuminant A.





elaborate and expensive, and there seems little prospect of improvement in this direction. The visual instrument is, in general, more robust and requires less maintenance. It is consequently suited to the unskilled or semi-skilled observer who, once the routine has been learned, can carry on with little skilled supervision.

There is another natural advantage of visual observation which should be mentioned, for it is missed very much when going over to photo-electric methods, that is, its sensitivity to low brightnesses. A visual instrument can be made to measure practically anything that can be seen. In addition to the measurement of filters and reflecting colours with the standard illuminants A, B and C, it can deal conveniently with any kind of illuminant and coloured specimens under that illuminant. The measurements can be made at ordinary levels of illumination, and it is not necessary to arrange for artificially high values of illumination to get the desired accuracy. When using direct observation of a tungsten filament it is possible to measure even the densest welders' protective goggles, which can reach an optical density of 6. It is for such reasons that visual methods cannot yet be regarded as superseded, but still have an important part to play in colour measurement.

In the following design of visual colorimeter two of the main defects of the ordinary trichromatic colorimeter have been removed: firstly, the large personal error of the observer, and secondly, the lack of sensitivity due to the rather small field. These improvements have introduced a little more complication in the instrument itself, and also a longer calculation in transforming the results. The observational work is reduced, however, so that, as far as the time for a complete measurement is concerned, there is an even balance between the two types of instrument.

§2. PERSONAL ERROR OF OBSERVER

This instrument can be regarded as an extension of the ordinary sphere colorimeter (Donaldson, 1935) with three mixing colours. As is well known, everyone can get a perfect colour match with three colours, but the settings vary with the observer. From the point of view of measurement this is a serious defect. Two different observers can, in measuring certain colours, get widely different results, although all their observations are closely grouped about their respective mean values. The cause of this is the combination of the observer's personal colour-vision characteristics and the differences of spectral energy distribution between the colours being matched. The colour being measured has in general a continuous distribution, whereas the instrument colour is a mixture of red, green and blue spectral bands only. In figure 1 are shown the spectral transmissions of the trichromatic instrument filters, which, when illuminated by illuminant A, form the instrument stimuli. It can be seen from figure 1 that there are big gaps in the energy distribution of the instrument colour.

In the present instrument three more mixing colours have been added, so that these gaps are filled in and the instrument colour is made to resemble more closely the colour being measured. A blue-green is inserted between the blue and the green, and a yellow-green and orange between the red and the green, this being the bigger gap. The filters were chosen to fit into each other as smoothly as possible so that the fall in transmission on one side of a filter is counterbalanced by the rise in an adjacent filter. In figure 2 are shown the spectral transmissions of the six filters and also the kind of fit with illuminant A when they are additively mixed together. Figure 3 shows the fit with illuminant B. For other smooth distributions, a similar order of fit is obtained.

Details of the construction of the filters

Red: Chance OR 1, 1·8 mm., and Calorex, 3·3 mm.
Orange: Chance OR2, 2·5 mm., Corning 978, 2·9 mm., and cadmium yellow, 0·9 mm.
Yellow-green: Chance OGr 1, 2·8 mm., and cadmium yellows, 1·7 mm. and 1·4 mm.
Green: Chance OY 4, 2·1 mm., and Zeiss BG 7, 4·2 mm.
Blue-green: Wratten gelatine filter No. 75.
Blue: Chance OB 1, 2·5 mm.

The cadmium yellows are unlisted yellow glasses in common use for signal glasses and fog lamps etc. They can be duplicated from Chance's later catalogues. It was found impossible to construct a suitable glass filter for the blue-green stimulus.

With three more mixing colours, there are six controls on the instrument and consequently there is no longer a unique setting for each colour match. The question therefore arises as to how the controls should be set so that there is an approximate energy match to the unknown colour that is being measured. The procedure for this is as follows. To set the red stimulus, a red filter of the same nature as the instrument filter is held at the eye and the red control is varied until there is a brightness match in the field. The process is repeated with an orange filter and the orange stimulus, and so on for each of the filters in turn. Owing to the slight overlap of the filters we require to repeat the process a second time, but very soon a state is reached where the instrument colour is in agreement with the colour being measured when viewed through each of the six filters in turn.

When the controls have been set in this way, there is in general an approximate but not an exact colour match in the field. To get an exact colour match three controls only—red, green and blue or orange, green and blue—are adjusted in the usual way. The amount of adjustment is so small that it does not disturb the energy match appreciably. At the matching point, therefore, the observer is only asked to discriminate between two colours of nearly the same energy distribution. Under these conditions there are no wide differences in the settings with different observers, and consequently the personal error is considerably reduced.

§ 3. CONSTRUCTIONAL.

The mechanical construction follows very closely that of the earlier sphere colorimeter (Donaldson, 1935). The linear scales of the stimuli are produced by apertures with sliding shutters and the colour mixing is carried out by an integrating sphere as in that instrument. The accompanying photograph, figure 4, shows the arrangement in the interior of the six-stimuli colorimeter, lamp A, condensing lens and apertures B, integrating sphere C and the photometric cube D. There are six rectangular apertures in front of the condensing lens. They are arranged in two columns, three on each side, with the sliding shutters opening outwards from the centre. The scales are engraved on the inside of the shutters and read by means of lenses and mirrors via the interior of the instrument. The filters are mounted on the outside. This is preferable to having them behind the shutters because in the outside position they are uniformly heated by the lamp.

It is important to have a smooth motion on the shutters without backlash. Transmission cables in tubes were used for three of the controls but, as they are not quite successful, pulleys and strings were fitted to the other three. The latter have proved to be quite satisfactory and provide a movement that feels pleasantly smooth and direct.

§4. FIELD SIZE

The removal of the energy differences between the colours being matched allows complete freedom in the choice of field size. In the trichromatic colorimeter the 2° field is standard. This size was adopted to ensure freedom from Purkinje effect over a large range of brightness and also to be in agreement with the standard observing conditions under which the response curves describing the colour and luminosity functions of the normal observer have been obtained. The practical need for this restriction only arises when there are appreciable differences in the energy distributions of the matched colours.

When the energy differences are removed or partially removed, as in this instrument, there is no need to restrict the size of the field. Large fields are more sensitive to colour differences than the small 2° field. A field of 15° angular size has therefore been chosen and the Lummer-Brodhun contrast patches have also been added. The Lummer-Brodhun field allows the eye to attain practically its limit of colour sensitivity. Small colour-differences which can be just seen under ordinary viewing always seem to be enhanced in the Lummer-Brodhun field. In colour-temperature work an accuracy of $\pm \frac{1}{2}\%$ in volts can easily be obtained with it. This corresponds to a maximum change of about 0·0005 in the trichromatic coefficients. The equivalent of this high discrimination is probably maintained over the whole of the colour field. As a result it is almost impossible to get a colour match on the instrument that is completely satisfying when looked at critically. Sufficient accuracy for all practical purposes can be obtained, however, by three or four quite casual matches. Matching casually saves a great deal of eye-strain. It is found with the majority of ordinary specimens that the variations due to non-uniformity generally tend to be greater than the smallest differences discernible in the field, so that as far as colour sensitivity is concerned the Lummer-Brodhun field is adequate.

§5. TRANSFORMATION EQUATIONS

The results as given by the instrument are arbitrary readings, in terms of scale divisions of red, orange, etc. A set of equations is therefore required which will transform to the C.I.E. standard reference stimuli X, Y and Z. In deriving the equations, two aspects of each mixing colour have to be defined. There is the colour quality, or *chromaticity*, and the amount that is present in the mixture. The colour quality is found by the usual method of spectrophotometry and calculation. The quantities of red, orange, etc., which correspond to a scale division cannot be obtained so directly. In the ordinary trichromatic colorimeter, these quantities are defined by means of a colour match made with white, usually standard illuminant B. With six colours there are too many to be related to each other by colour matching, but they can be related by a series of brightness matches. The auxiliary filters used to analyse the spectral distribution of the colour being measured also serve as standards for the brightness matches. In this measurement they are not placed at the eye but in the usual position for the measurement of transparent specimens and illuminated by illuminant A. They are of the same colour and energy distribution as the respective instrument stimuli so that the results are independent of the observer's colour vision. The transmission of each auxiliary filter is known, so the quantities of the matching stimuli can be related to each other and the transformation equations derived. The transformation equations are six in number, and a typical example is as follows :

$$\begin{aligned} R &= 1{\cdot}847X+0{\cdot}696Y+0{\cdot}000Z \\ O &= 11{\cdot}387X+5{\cdot}933Y+0{\cdot}005Z \\ YG &= 5{\cdot}943X+8{\cdot}031Y+0{\cdot}073Z \\ G &= 0{\cdot}730X+3{\cdot}360Y+0{\cdot}435Z \\ BG &= 0{\cdot}200X+0{\cdot}849Y+1{\cdot}652Z \\ B &= 1{\cdot}167X+0{\cdot}197Y+6{\cdot}231Z \end{aligned}$$

The quantities R, O, etc., refer to one unit of each of the matching stimuli, one scale division of red, orange, etc. The right-hand side of each equation is proportional to the trichromatic coefficients defining the chromaticity of the instrument stimulus. The proportions are such that the ratios of the coefficients of Y are as the relative luminosities of one division of red, orange, etc. The method of using the equations is the same as that for the ordinary trichromatic transformation.

There is one important difference, however, between sets of equations derived in this way and by the method used in the trichromatic colorimeter, i.e. the white is no longer automatically given the correct value. The measurement of white is the same as for any other colour, and small experimental errors may appear in it. When measuring colours close to white, we can take advantage of the readily available standard, magnesium oxide, and use the instrument as a differential colorimeter to measure the difference between the standard and the near-white. This difference, if small, will be free from any systematic error due to the instrument.

§ 6. ACCURACY

The ideal measuring instrument should give results conforming to the average observer when used by ordinary observers having the usual variations in colour vision. These variations should not be capable of seriously upsetting the instrumental results. As the spectral energy matches in this instrument are never quite exact, the residual differences may cause some slight variation with observer. There are also the experimental errors in the determination of the constants of the instrument, e.g. in the colours of the filters, in the brightness matches required for the derivation of the transformation equations and in the setting of the templates controlling the linearity of the matching stimuli. All these factors taken together seem to have a greater influence on the experimental error than the chance variations in matching, which are very low on account of the high sensitivity of the Lummer-Brodhun field. This is shown by the fact that the repetition by a single observer is often better than the agreement with the normal observer.

It is also noticed sometimes that observers show a bias in a given direction with certain colours. This would seem to indicate that colour-vision variations are no longer the most important factor reducing the accuracy. There are the small, residual errors, due to the system of shutters and templates, and (what is probably more important) those due to the use of glass filters. Filters are never strictly uniform. They show variation in colour over their surface, and it needs very careful selection to keep this down to negligible proportions. Our experience of the instrument has shown that for the majority of colours the personal error of the observer has been reduced to the order of those inevitable errors arising from the mechanical construction of the instrument.

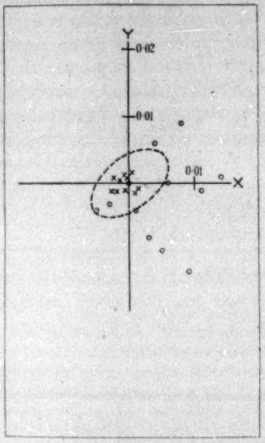


Figure 5. Errors in colour-measurement.

Yellow

$0{\cdot}6000X + 0{\cdot}3993Y + 0{\cdot}0007Z$

× 6-stimuli instrument.
⊙ 3-stimuli instrument.
- - - - MacAdam ellipse.

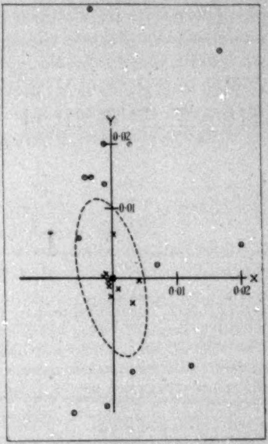


Figure 6. Errors in colour-measurement.

Blue-green

$0{\cdot}1918X + 0{\cdot}3976Y + 0{\cdot}4106Z$

× 6-stimuli instrument.
⊙ 3-stimuli instrument.
- - - - MacAdam ellipse.

In figures 5 and 6 are shown comparisons with this instrument and the ordinary trichromatic colorimeter of measurements made with two coloured filters, a yellow and a blue-green. These results refer to three observers and have been obtained at various times in the course of testing trichromatic colorimeters and the measurement of signal colours on the six-stimuli instrument. The origin of co-ordinates is the calculated colour of the filter. To give some indication of the colour sensitivity in these regions of the colour chart the MacAdam (1942) ellipses on a scale of three times the standard deviation have been sketched in. The ellipses represent a just noticeable colour difference. The improvement of the six-colour instrument over the three is very marked for these colours.

This instrument has been in service for some years now and it has been found that in general its accuracy with other colours is of the same order as shown on the diagrams.

§7. ACKNOWLEDGMENTS

The development of this instrument has been carried out as part of the research programme of the National Physical Laboratory, and this paper is published by permission of the Director of the Laboratory.

REFERENCES

Donaldson, R., 1935. *Proc. Phys. Soc.*, **47**, 1068.
Knipe, G. F. G. and Reid, J. B., 1943. *Proc. Phys. Soc.*, **55**, 81.
MacAdam, D. L., 1942. *J. Opt. Soc. Amer.*, **32**, 247.
Winch, G. T., 1946. *Trans. Illum. Engng. Soc., Lond.*, **11**, 107.

THE RECOGNITION OF COLOURED LIGHT SIGNALS WHICH ARE NEAR THE LIMIT OF VISIBILITY

By N. E. G. HILL,
Royal Aircraft Establishment, Farnborough, Hants

MS. received 24 *September* 1946

ABSTRACT. Statistical tests on the recognition of colour were made during 1938–39 to find the range of colours which would be best for aviation signals. Seventy-three colours were seen as point sources and viewed by binocular foveal vision, with dark-adapted eyes against a dark background, by nine observers of normal colour vision. The results were plotted as recognition contours, for eye illuminations of 1 mile-candle and 2 mile-candles respectively, on the x, y colour diagram for the colour categories red, yellow + orange, green + blue, and white. The results indicate that yellow + orange is the least satisfactory colour group for signals of low illumination. A modification is suggested to the specification for " aviation white " to avoid the risk of confusion with yellow + orange.

§1. INTRODUCTION

A characteristic feature of a coloured light signal is that its colour becomes less pronounced as the illumination of the signal at the observer's eye is reduced, and may disappear entirely before the limit of visibility is reached, so that at low values of illumination the chance of confusion between colours is increased. This effect is more marked with the paler or less saturated colours. In choosing colours for long-range light signals it is therefore necessary to select those colours which are the most recognizable when seen as point sources of low illumination. The choice of coloured signals has always been based on accumulated experience with particular colours, but it was thought, during the years before the war, that systematic data should be obtained on the recognition of coloured point sources in order that the full range of possible colours might be known. Data of this kind were obtained at the Royal Aircraft

Establishment during 1938 and early 1939 in an endeavour to find the best colours for aviation signals but, owing to the war- time restrictions, these data could not be published until now. Thus, although the data are not recent, they are new in the sense that they have not previously been published, and they are presented in the present paper in the belief that they may still serve as a contribution to the knowledge of colour recognition.

The tests here described were made with binocular foveal vision under conditions closely related to those under which aviation signals are usually observed but, in order to obtain consistent data which could be compared with those of other investigators, the conditions were idealized and closely controlled.

The background brightness was fixed at about that of starlit sky, the effects of atmospheric absorption were eliminated, and the tests were confined to the five groups of colour which are usually used for aviation signals, viz., red, yellow and orange, white, green, blue. Of these colours, blue is normally used only as a short-range signal, but the remainder are long-range colours. It was decided not to attempt to separate yellow recognition from orange because it seemed unlikely that this could be done satisfactorily at low values of eye illumination. Nor was any attempt made to obtain recognition figures for purple, which is known to be unsatisfactory at long range.

§ 2. THE PROBLEM

The recognition of a coloured light signal is a subjective reaction which in general cannot be predicted absolutely for a single observation, even for the average observer, but which can, however, be expressed as a probability for a single observation. For instance a particular signal, seen 100 times by an average observer under certain conditions specified, might be judged to be red 85 times, yellow 10 times, and white 5 times. The particular colour would therefore be recognized as red on 85% of the observations, and would have an 85% probability of being recognized as red on any single observation. Such a colour may be defined as having a red recognition of 0·85 under the specified conditions of observation. The basic problem in obtaining recognition data is thus a statistical one, and it was this consideration which governed the arrangements for the tests here described.

It was evident that a large number of observations of each signal would be required, and that each observation must be an independent one, unprejudiced by the judgment made on any previous observations of the same signal. The best solution to this problem appeared to be to present a succession of coloured signals, in random sequence, to an observer who was required to place each colour in one of a number of specific colour categories, and to repeat the process until sufficient observations were made. This method is similar to that used by McNicholas (1936) in a series of tests on signal glasses to determine the best set of six colours for use in a system of railway signals.

Some additional requirements had considerable influence on the apparatus and methods used. These were that the colours of the test signals should be spread as widely as possible over the colour diagram, that a number of normal observers should participate to an equal extent, that each signal should be observed a large number of times by each observer, and that random errors should be reduced by strict control of test conditions.

§3. THE COLOURED SIGNALS

The production of a large number of coloured point-signals, adequately spread over the possible range of colours, presented difficulties. A sufficient number of different single filters was not available, and the scheme of mixing coloured lights in various proportions was therefore considered.

The principle of the trichromatic colorimeter offered attractive possibilities both for the production of an adequate range of colours and for simplicity of control. The theoretical and experimental bases of the trichromatic theory have been clearly and adequately stated by Judd (1930) and Guild (1931) and summarized by Stiles, Bennett and Green (1937) and others, and are too well known to need further discussion here.

Three filters were chosen having, in conjunction with a standard illuminant, colours R, G, B, spaced widely over the colour diagram in figure 1. Light from these three filters was mixed in a diffusing sphere having a window covered by a pinhole. The quantity of light through each filter was controlled by means of a shutter and the colour and intensity of the illumination on the pinhole could thus be varied within wide limits. It will be clear from figure 1 that any colour C within the triangle RGB could be produced at the pinhole.

It was however found that this method of producing coloured point sources failed because, due to chromatic aberration in the eye, the component colours of the mixture were separated, and the apparent colour of the point was entirely changed. For instance, if the proportions of the primary colours were arranged to give what appeared a good white when the sphere window was viewed at short range, then, at long range, the pinhole appeared as a red point surrounded by a green-blue halo. It is clear that the light produced by mixing the three colours in the way described has a spectral composition which may be entirely different from that of light from a single filter of the same colour as the mixture. It was thought that satisfactory results would be obtained only if the spectral composition of the light used for the coloured point source was similar to that of the light from a single filter and light-source combination of the same colour.

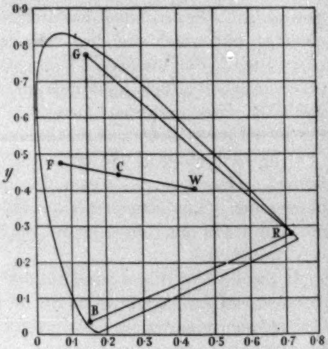


Figure 1. Colour diagram on 1931 I.C.I. standard reference system, showing method of obtaining colour mixtures.

The trichromatic method of producing coloured point sources was therefore abandoned and the alternative hue-and-saturation method was tried. A number of filters was obtained whose colours, in conjunction with a filament lamp, were represented by points as near the boundaries of the colour diagram as possible. Light from any one of these filters was mixed with white light in the diffusing sphere already referred to, and the mixture used to illuminate a pinhole. By varying the proportions of colour F and white W, a large range of colour mixtures C was made available (see figure 1).

The spectral distributions of such mixtures are similar to, though not identical with, the distribution of light from a single filter of the type encountered in practice. No difficulty due to eye aberration was experienced in viewing point sources formed in this way, and it was decided that, since the spectral transmission of the filters used was specified (1938), and since the spectral energy distribution of the point sources could thus be calculated if desired, recognition tests could usefully be made.

§ 4. DESCRIPTION OF APPARATUS

It was clear that, as a very large number of separate observations would be required, great care would have to be taken to ensure consistent reproduction of each signal and to avoid tiring the observers. The test apparatus was therefore designed with a view to ensuring the accurate presentation of each colour, and rapid change from colour to colour.

The apparatus is represented in diagrammatic form in figure 2. The coloured filters were mounted on a vertical disc (2), and light from the lantern (1) passed through a filter and the clear glass sheet (3) into the diffusing sphere (5). The light source consisted of a 200-v., 500-w. class A1 projector lamp backed by a plane silvered-glass mirror. The filter disc was arranged to rotate and was provided with a ratchet so that any desired filter could be quickly and accurately brought into position. The "white light" source (4) consisted of a 12-v., 60-w. motor-car type lamp, also backed by a plane silvered-glass mirror. The light from this source was reflected from the clear glass sheet into the diffusing sphere for mixing with the coloured light. Each lantern was arranged on slides along its light axis and was provided with an index and calibration scale, the two sets of slides being at right angles.

The details of the diffusing sphere are shown in figure 3. The light entered the sphere through the larger opening and was prevented from passing right through by one central flat screen. The inside of the sphere and the screen were silver plated, polished, and then coated with a uniform layer of magnesium oxide. The coloured and white lights were completely mixed inside the sphere, and the composite light emerged from the smaller opening of the sphere, outside which was placed a pinhole of 0·0496 inch diameter.

The pinhole, which acted as a luminous coloured point source, was viewed by the observer seated at 24·5 feet distance. The position of the observer's eyes was fixed by a binocular eyepiece which could be adjusted to suit the distance between the eyes, and which did not restrict the pupil.

The wide range of transmissions of the filters used for the tests necessitated the provision of a variable sector disc (7), figure 2. The point source was made visible to the observer by means of a rotary shutter on the flasher unit (8), which was controlled electrically and arranged to give a single flash of definite duration.

It was found desirable to provide a means of focusing the observer's eyes on the correct spot before the flash of the point source occurred. If this was not done, a considerable part of the flash was wasted while the eye searched for the point and then focused on it. Accordingly light from a 12-v., 4-w. lamp (10), operating at 8 volts, was allowed to fall on the black surface of the flash shutter for about 1½ seconds immediately prior to the flash. This preliminary light appeared to the observer as a dim area of illumination covering the aperture in the screen (9); the angular diameter of the area was about 15 minutes of arc, and its brightness of the order of 0·0005 candles per square foot; this was found to be very suitable for the purpose and formed an essential part of the test apparatus. The preliminary light, because of its low brightness and of its dissimilarity in character from the main signal, did not prejudice the observer in his opinion of the colour of the point source. As an additional help to the observer, a warning gong was sounded once about ½ second before the preliminary light appeared. The sequence was initiated by a push button and operated through a system of relays controlled by cam contacts on the flasher motor.

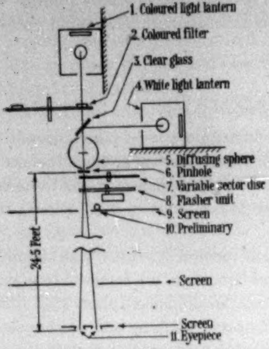

1. Coloured light lantern
2. Coloured filter
3. Clear glass
4. White light lantern
5. Diffusing sphere
6. Pinhole
7. Variable sector disc
8. Flasher unit
9. Screen
10. Preliminary
24·5 Feet
Screen
Screen
11. Eyepiece

Figure 2. Arrangement of apparatus for colour-recognition tests.

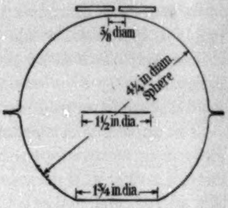

⅜ diam
4½ in. diam. sphere
1½ in. dia.
1¾ in. dia.

Figure 3. Diffusing sphere.

The tests were made in a photometric dark room, the black walls of which formed the general background of vision. Screens were placed as shown in figure 2 to intercept stray light from the test apparatus.

§ 5. CALIBRATION OF APPARATUS

Wratten light filters were used, consisting of 2-inch squares of coloured gelatine sandwiched between two clear-glass plates. During the tests there was no spectrophotometer available, and the properties of the filters were therefore calculated from the wavelength-transmission data published by the Kodak Co. (1938) and, in addition, were measured by visual photometry.

The total transmissions of the filters were measured by a flicker photometer, and the values were checked by comparison with standard filters on a Lummer-Brodhun contrast head. The colour coefficients were measured by means of a

Donaldson colorimeter. All these measurements were made by several observers whose results were averaged.

More recently the wavelength-transmission characteristics of the filters were measured by means of a photoelectric spectrophotometer, with the exception of No. 22, which had been so measured for threshold experiments soon after the conclusion of the present tests, and No. 23, which was no longer available. The measured spectral transmission values are given in table 1, together with certain additional data from the Kodak specification. From these measured values the total transmission and the colour coefficients were calculated for each filter for a 2848° K. source.

Table 2 gives a comparison of the filter properties obtained (*a*) by calculation from the Kodak specification, (*b*) by calculation from the spectrophotometric measurements, and (*c*) from the visual photometry. It will be seen that a very fair measure of agreement on colour exists among the three sets of data and, for most of the filters, the x, y coordinates of each set lie within $\pm 0{\cdot}005$ of the mean. The chief exceptions are filters 34, 47 and 63, whose precise colours are therefore in doubt.

In view of this general agreement, and in spite of the lapse of time, it seems reasonable to assume that the measured spectrophotometric data are a fair representation of the filters at the time of the tests. The x, y coefficients given under (*b*) in table 2 have therefore been used to plot the test results on the colour diagrams except in the case of filters 23, 33, 45 and 73. For these four filters the coefficients given under (*a*) have been used because they agree more closely with the visual measurements under (*c*). For the sake of completeness the Kodak data for these four filters are included in table. 1

In the absence of other data the transmission values obtained by visual photometry were used to calculate the lantern adjustments, etc., for each particular signal (see §7), but in any case the visual data, which were obtained by careful measurements, may be regarded as the most reliable assessment of the transmission values at the time of the recognition tests.

The 500-watt and 60-watt filament lamps, used as light sources in the coloured and white light lanterns respectively, were calibrated for 2848° K. colour temperature by matching with an N.P.L. standard lamp on a Lummer-Brodhun head. The lamps for the transmission and colour measurements were similarly calibrated.

The scale of each lantern was calibrated for white light in terms of the illumination at the observer's eyes. To do this the pinhole was removed and the brightness of the output window of the diffusing sphere was measured, for various positions of the lantern, by means of an illumination photometer. It was calculated, from the diameter of the pinhole and the distance of the observer's eyes, that an eye illumination of 1 mile-candle would require a brightness of 1·60 candles per square foot in the sphere window; thus a calibration curve of eye illumination against lantern position was obtained. The illumination photometer was itself calibrated by an N.P.L. standard candle-power lamp on a photometric bench.

Both the colour temperature and the illumination scale of the lamps were checked from time to time during the course of the tests. The diffusing sphere was twice cleaned and re-coated with magnesium oxide, and it was also found necessary to renew the silvered-glass mirror in the coloured-light lantern.

Table 1. Measured spectral transmission data for Wratten filters

K, data from Kodak specification. R, measured in 1940. Other data measured in 1946. Transmission in % at each wavelength.

λ (mμ)	15	22 R	23 K	24	31	32	33	33 K	34	45	45 K	46	47	47*a*	63	65*a*	71*a*	73	73 K	74
400	—	—	—	—	14·5	35·5	1·5	1·8	60·6	—	—	3·3	22·3	9·7	—	—	—	—	—	—
410	—	—	—	—	16·5	35·3	1·6	1·3	64·7	—	—	2·0	27·2	21·5	—	—	—	—	—	—
420	—	—	—	—	19·2	37·7	1·8	1·4	66·0	—	—	3·2	34·7	30·5	—	—	—	—	—	—
430	—	—	—	—	25·0	42·3	2·3	2·2	63·3	1·2	2·7	12·2	41·0	37·0	—	2·5	—	—	—	—
440	—	—	—	—	41·7	50·2	4·8	10·0	58·3	10·7	18·2	16·7	46·0	39·5	—	8·7	—	—	—	—
450	—	—	—	—	52·1	60·5	14·3	15·8	50·0	28·5	27·6	29·0	49·2	38·2	—	17·5	—	—	—	—
460	—	—	—	—	45·7	61·7	16·8	10·0	37·2	41·0	34·7	36·0	48·3	34·0	—	27·5	—	—	—	—
470	—	—	—	—	27·6	52·5	5·5	3·1	23·6	46·4	39·9	37·0	42·5	27·0	1·3	37·5	—	—	—	—
480	—	—	—	—	11·3	37·7	—	0·1	11·2	47·8	41·5	31·7	35·0	19·7	3·7	46·5	—	—	—	—
490	—	—	—	—	3·7	22·5	—	—	3·3	45·3	39·9	23·4	26·5	12·0	7·5	53·0	—	—	—	—
500	—	—	—	—	0·8	11·2	—	—	0·5	39·3	34·9	14·5	18·5	5·6	11·2	53·5	—	—	—	—
510	0·5	—	—	—	—	4·0	—	—	—	29·5	28·3	6·3	10·7	1·6	14·0	49·5	—	—	—	—
520	15·5	—	—	—	—	1·0	—	—	—	17·2	16·9	1·5	4·3	—	14·8	39·8	—	—	—	13·2
530	57·0	—	—	—	—	0·5	—	—	—	7·0	8·0	0·3	1·5	—	12·8	27·5	—	—	—	10·0
540	79·5	—	—	—	—	0·5	—	—	—	2·0	2·4	—	0·5	—	9·5	15·3	—	—	—	8·7
550	87·2	0·5	—	—	—	0·5	—	—	—	0·5	0·1	—	—	—	5·5	7·5	—	0·2	—	4·1
560	89·7	28·2	—	—	—	0·5	—	—	—	—	—	—	—	—	2·3	2·5	—	7·3	2·5	1·3
570	90·5	70·5	2·5	—	—	0·5	—	—	—	—	—	—	—	—	0·5	0·8	—	16·2	8·0	0·2
580	90·5	83·0	34·7	2·0	—	0·5	—	—	—	—	—	—	—	—	—	0·2	—	11·0	5·7	—
590	90·5	85·7	66·5	32·5	3·6	0·5	—	—	—	—	—	—	0·2	—	—	—	—	4·9	2·7	—
600	90·5	86·8	76·0	72·4	42·0	12·7	—	—	—	—	—	—	1·8	—	—	—	—	2·2	1·2	—
610	90·5	87·0	79·8	83·5	75·7	57·7	2·5	3·0	—	—	—	—	2·4	—	—	—	0·5	1·0	0·4	—
620	90·5	87·3	82·0	86·6	86·0	80·3	36·0	39·6	—	—	—	—	1·7	—	—	—	4·2	0·5	0·2	—
630	90·5	87·6	83·6	87·5	88·7	86·0	70·0	67·5	0·2	—	—	—	1·2	—	—	—	6·6	0·2	—	—
640	90·5	87·9	85·0	87·6	89·8	87·5	82·7	80·0	4·0	—	—	—	0·8	—	—	—	7·1	—	—	—
650	90·5	88·2	86·2	87·8	90·0	88·0	86·5	82·5	20·0	—	—	—	0·5	—	—	—	6·8	—	—	—
660	90·5	88·4	87·0	87·9	90·0	88·3	88·0	84·5	44·0	—	—	—	0·5	—	—	—	6·2	—	—	—
670	90·5	88·7	87·5	88·1	90·0	88·3	88·0	85·5	63·5	—	—	—	0·5	—	—	—	5·5	—	—	—
680	90·5	89·0	87·7	88·2	90·0	88·3	88·0	86·5	75·0	—	—	—	0·5	—	—	0·9	5·0	0·8	0·2	—
690	90·5	89·3	88·0	88·4	90·0	88·3	88·0	86·8	81·0	—	—	—	0·5	—	—	4·3	5·4	2·8	2·2	—
700	90·5	89·6	88·0	88·5	90·0	88·3	88·0	87·0	84·5	—	—	—	0·5	—	—	6·5	6·4	5·8	6·3	—

Table 2. Colour and transmission of Wratten filters with 2848° K. source

Wratten filter number	(a) Kodak specification			(b) Spectrophotometry			(c) Visual photometry		
	Colour		Transmission (%)	Colour		Transmission (%)	Colour		Transmission (%)
	x	*y*		*x*	*y*		*x*	*y*	
15	0·548	0·452	74·9	0·542	0·454	79·0	0·546	0·450	77·3
22	0·623	0·377	46·2	0·615	0·385	49·7	0·620	0·380	49·1
23	0·654	0·346	32·5	—	—	—	0·653	0·347	36·7
24	0·675	0·325	25·0	0·675	0·325	26·8	0·673	0·321	26·6
31	0·609	0·263	19·4	0·606	0·267	22·0	0·607	0·272	20·9
32	0·543	0·241	20·2	0·553	0·243	19·0	0·551	0·240	19·7
33	0·675	0·267	9·90	0·686	0·254	10·0	0·675	0·269	9·6
34	0·253	0·064	0·97	0·316	0·095	2·00	0·353	0·115	1·90
45	0·105	0·258	4·00	0·104	0·250	4·13	0·112	0·258	3·3
46	0·121	0·109	1·06	0·122	0·109	1·24	0·123	0·110	1·05
47	0·138	0·070	1·13	0·170	0·118	2·32	0·184	0·131	2·1
47*a*	0·145	0·043	0·46	0·142	0·052	0·65	0·134	0·058	0·52
63	0·120	0·670	3·54	0·159	0·689	3·71	0·152	0·693	3·1
65*a*	0·110	0·437	7·12	0·118	0·436	9·28	0·121	0·436	8·0
71*a*	0·709	0·291	1·29	0·710	0·290	0·90	0·712	0·288	1·31
73	0·496	0·503	1·87	0·490	0·510	3·93	0·500	0·497	4·25
74	0·200	0·766	2·62	0·225	0·750	1·92	0·232	0·745	1·9

§6. CONDITIONS OF EXPERIMENT

When giving the results of photometric, colorimetric and other tests involving visual observation, it is desirable to state the precise conditions under which the observations were made. The conditions under which the recognition measurements were made are therefore summarized here.

Two series of recognition tests were conducted, the first at an eye-illumination of 1 mile-candle and the second at an eye-illumination of 2 mile-candles. The point sources were viewed by binocular foveal vision, with dark-adapted eyes, against a dark background for a period of $1\frac{1}{2}$ seconds. The angular diameter of the point source was 0·6 minutes of arc. The general background brightness was about 0·0001 candles per square foot, or of the order of brightness of a starlit sky.

The various colours were shown in succession in random sequence; there was an approximately equal number of each class of colour so that no class was unduly emphasized. During the tests each colour was seen alone and could not be contrasted with any other light. The tests were performed by nine male observers who were considered, from the results of certain transmission measurements and colorimeter tests, to have normal colour vision. The observers were tested by the Ishihara colour charts and all classed as normal. The age groups of the observers are shown in table 3.

Table 3. Age groups of observers

Age :	20–24	25–29	30–34	35–40	>40	Average :	30
Number :	3	2	2	1	1	Total :	9

§7. EXPERIMENTAL PROCEDURE

Seventeen coloured filters were used, and from these, by the addition of white, a total of 73 colours was produced. Each colour was identified by the number of the Wratten filter followed by a letter representing the degree of relative saturation of the colour as compared with the pure filter colour. The settings of the two lanterns and the adjustment of the variable sector disc were calculated for each colour. These settings, together with the identification number and letter of the colour, were written down on small index cards, there being one card to each colour and eye-illumination. The two series of 1 mile-candle and 2 mile-candles were taken separately and each series was divided into two groups to avoid tiring the observers. The cards in a group were shuffled before each test to preserve a random sequence.

The procedure in carrying out a test was as follows. The observer was allowed to get dark-adapted (usually about 10 minutes was sufficient for this), then, having adjusted his eyepiece, he observed each colour in turn. The observer was required to say, after each flash, what colour he considered the signal to be, the choice of colour being restricted to the following five categories: red, yellow or orange, white, green, blue. A definite decision was required on each observation, and no repetition of a colour was permitted unless the observer failed to see the signal because he blinked or was out of position. After each observation the operator readjusted the apparatus in accordance with the settings indicated on the next card. An assistant kept the voltage on the lamps at the correct settings, and also recorded the identification number and letter from each card together with the colour category named by the observer. The tests were repeated from day to day, over a period of nine months, until each colour had been seen by each observer 20 times, and in the case of the first group 30 times. The alternative designation of yellow or orange was permitted because some observers experienced a psychological reluctance to be limited to yellow. In assessing the results all yellow and orange designations were taken together.

§8. EXPERIMENTAL RESULTS

The observations recorded during the tests were sorted and tabulated for each colour. Table 4 shows this tabulation in the case of colour 31G. The successive readings of each observer are given, together with his total number of recognitions in each colour category. The results are added for all the observers and the percentage recognition evaluated. The symbol Y in the table includes both yellow and orange designations.

Table 4

Recognition record for colour	31 G
Eye illumination	2 mile-candles
Filter saturation	0·6

Observer	Recognition		R	Y	W	G	B
L. N. B.	R R R R Y R Y R Y R	R R Y R R R Y R Y Y	13	7	–	–	–
E. S. C.	Y R Y R Y W R R R W	W W W R Y W W Y Y Y	6	7	7	–	–
J. C. C.	R R Y R Y R R R R R	Y R R R R Y R Y Y R	14	6	–	–	–
S. H. G. C.	Y Y Y Y Y Y Y Y Y Y	Y Y Y Y Y Y Y Y Y Y	–	20	–	–	–
H. N. G.	Y R Y Y W Y Y Y Y R	Y R R Y R Y R R R R	9	10	1	–	–
N. E. G. H.	R Y Y Y R Y Y R R Y	Y Y R R Y R R Y R R	10	10	–	–	–
R. E. L.	R R R R R R R R R Y	R R R R R R R R R R	19	1	–	–	–
D. B. McK.	R R R R R R R R R R	R R R R R R R R R R	20	–	–	–	–
J. W. S.	R R R R Y R R R R R	R Y R R Y R R R R R	17	3	–	–	–
		Total	108	64	8	0	0
		Percentage	60	36	4	0	0

The percentage recognitions thus found were plotted as ordinates against the relative saturation, S, as abscissae, one set of curves being plotted for each filter at each eye illumination. Figure 4 shows the set of curves for filter No. 45 at 2 mile-candles illumination.

Interpolation on the above curves enabled recognition contours to be plotted on the colour diagram. In a few cases the highest recognition required a relative saturation greater than unity, and extrapolation of the curves was made in these cases; the resulting points on the contours lie between the pure filter points and the spectrum locus.

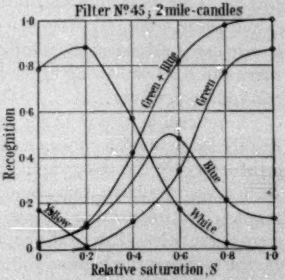


Figure 4. Typical saturation-recognition curves.

The method of calculating the colour coefficients corresponding to a given relative saturation of a filter is a particular case of the general problem of calculating colorimetric purity which has been analysed by Judd (1931). It can be shown that, referring to figure 1, if x_w, y_w, z_w are the coefficients of the white point W, x, y, z, are those of C, and x_f, y_f, z_f, those of F, then

$$x = P \, . \, x_f + (1-P)x_w \qquad \ldots\ldots(1)$$

and

$$y = P \, . \, y_f/S, \qquad \ldots\ldots(2)$$

where

$$1/P = 1 + (1/S - 1)y_f/y_w, \qquad \ldots\ldots(3)$$

and where S is the relative saturation of the filter as defined in § 7. The quantity P, given by equation (3), is the ratio of the distance CW to the distance FW, and may be termed the *relative excitation purity*.

§ 9. DISCUSSION OF RESULTS

The tests which have been described produced a total of 30,420 observations taken over a period of nine months. The results have been summarized in the form of colour-recognition contours plotted on the standard I.C.I. colour diagram in figures 5 and 6, for 1 and 2 mile-candles, respectively.

The data should be strictly applied to signals having the same spectral-energy distribution as those used for the experiments, but it seems improbable that recognition can be critically dependent on spectral-energy distribution, and such isolated rough checks as have been possible suggest that a wide variation of spectral distribution has a relatively small effect on recognition. It therefore seems reasonable to apply the present results to the coloured signals which are used in practice, but further experimental data are required to confirm the validity of this procedure.

It was found that green and blue signals could not be distinguished from one another with any certainty at the low illuminations used for the tests. It is clear that green and blue would not be suitable for use as two separate signal colours at long range. The recognition values of green and blue were therefore added to obtain recognition contours for a single signal colour called green + blue in figures 5 and 6. This procedure does not imply any new restriction on the choice of signal colours in practice, as blue signals (i.e. signals which appear blue at short range) are in any case unsatisfactory at long range because the eye has difficulty in focusing them and because, owing to the low luminosity of blue light, blue filters have low transmission values.

A noticeable feature of the results is that the points plotted in figures 5 and 6 readily form smooth contours for the green + blue and for the white, but are not satisfactory for the red group. In the case of the yellow + orange group the points were so scattered that it was thought best to draw the nearest smooth curve through them. The importance to be attached to the yellow + orange contours is therefore considerably less than to the contours of the other colours, and this is emphasized by the fact that the highest yellow recognition point available was 80% at 2 mile-candles and 70% at 1 mile-candle, as compared with 100% for green + blue and red. It is concluded that yellow + orange is the least satisfactory group for a signal colour at low illumination.

In view of the poor recognition of the yellow + orange group it would be expected that it would be very difficult indeed to obtain orange as a separate group. This was in fact the experience during the tests.

Lines representing the limits defined in B.S.S.563/1937 (see Appendix) for aviation colours have been drawn. It will be seen that the areas thus defined are areas of high recognition except in the case of white, which extends into the region of yellow recognition. It appears that the specification for aviation white was framed so as to include the paraffin flame, but there seems to be little justification for this and, in view of the danger of confusion with yellow, it is thought that the specification should be amended so that, instead of "x not greater than 0·540", it should read "x not greater than 0·477". Filament lamps operating at colour temperatures down to 2500° K., including all the lamps usually used for aviation purposes, would thus fall within the specification.

If it be assumed that 80% or higher recognition is satisfactory, there is a considerable area of satisfactory green + blue recognition outside the B.S. specification. This additional area is, however, not a useful area because of the practical objections to blue signals already mentioned. There is also a large area of high red recognition outside the specification, but if the specification were extended to cover this area, which lies in the blue direction, the short-range appearance of the colours might be unsatisfactory. It is clear that any extension of the red specification along the spectrum in the direction of shorter wave-lengths would be unsafe.

The data presented in the contours of figures 5 and 6 were obtained under conditions where atmospheric absorption has no appreciable effect on the results, and where no searching of the field of view was required. No precise data are yet available on the change of colour of light transmitted through hazy atmosphere, or on the influence on recognition of searching the field for the signal. The comparison with the B.S. specification for aviation colours has not therefore taken into account these two factors.

The effect of increasing the illumination at the eye from 1 to 2 mile-candles is not anywhere very great. The greatest effect is noticed in the green + blue region, where the recognition is raised about 10%. The deep blue, white (on the purple side), and red (except spectrum reds) are unaffected. It seems likely that further increase of illumination would not give a proportionate increase of recognition except possibly in the case of the yellow + orange group. It also appears that small errors in the adjustment of the intensity of the signals will not have had an important effect on the results.

All the purples used in the tests appeared red when seen as point sources, the usual dichromatic characteristic of purple point sources being absent.

Tests were made to determine whether a flash period longer than 1½ seconds would affect the results. No change in recognition could be detected. The recorded results were studied closely to discover whether the results of test depended on the order in which the observer viewed the colours. No such dependence could be detected, and it was concluded that any given observation was unaffected by the previous observation.

It would naturally be of the greatest interest to compare the data of the present paper with the results obtained by other investigators. There are two other sets of published results which were obtained by somewhat similar test methods and which might be taken for comparison. The recognition tests of McNicholas have already been referred to. Another series of tests was made at about the same time as the present tests and subsequently published by Holmes (1941).

The results obtained by McNicholas in his tests unfortunately suffer from two

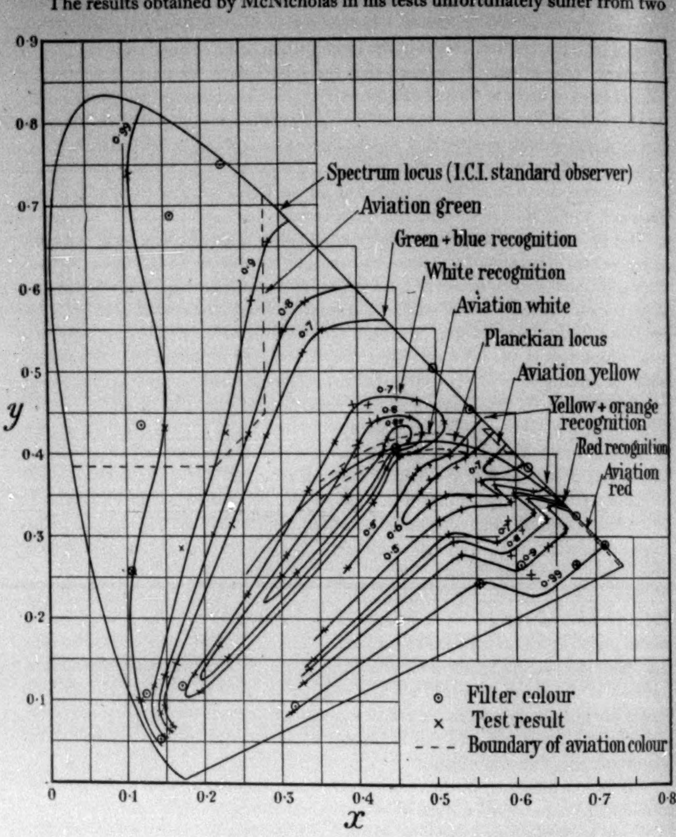


Figure 5. Colour-recognition contours for 1 mile-candle point sources.

limitations: first, the results represent the average recognition of signals whose illuminations range from 0·40 to 6·2 mile-candles in one series of tests, and between still wider limits for other series; and second, only a single line is given for each colour category instead of an area of recognition as in the present tests. It is therefore difficult to compare the results of the present tests with McNicholas's results. It is, however, of interest to note that he finds that green and blue are not easily distinguished, a conclusion which is in agreement with the results of the present tests.

Holmes's tests were made with an apparatus of excellent design, and had the advantage that 256 coloured signals were used. Unfortunately these tests also suffer from a severe limitation; it is that each signal was seen no more than three

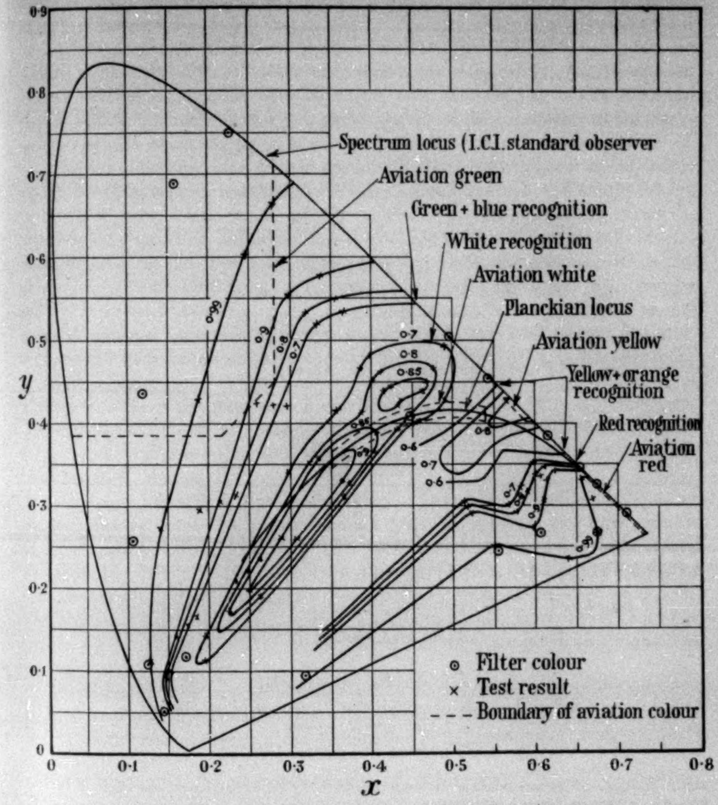


Figure 6. Colour recognition contours for 2 mile-candles point sources.

times by each of six observers. Since most of the signals were likely to be recognized, at least occasionally, in three or more of the colour categories, it is clear that three observations per observer are quite insufficient to yield statistically significant results. The large number of signals used does not compensate for this deficiency. It is therefore not surprising that, while a general similarity with the present results exists, there is no detailed concordance.

§ 10. ACKNOWLEDGMENTS

The author wishes to express his gratitude to colleagues who acted as observers during the tests. He also wishes to thank the Chief Scientist, Ministry of Supply, for permission to publish this paper.

APPENDIX

Aviation colours (B.S.S.563/1937. Appendix A)

	x	y	z
Aviation red	—	$\leqslant 0{\cdot}335$	$\leqslant 0{\cdot}002$
Aviation yellow	—	$0{\cdot}402$ to $0{\cdot}430$	$\leqslant 0{\cdot}007$
Aviation green	$\leqslant 0{\cdot}280$ $\leqslant y-0{\cdot}170$	$\geqslant 0{\cdot}385$	—
Aviation white	$0{\cdot}350$ to $0{\cdot}540$*	—	—

* $|x-y_0| < 0{\cdot}01$, where y_0 is the y-coordinate of the Planckian radiator for which $x_0=x$.

REFERENCES

Guild, J., 1931. *Phil. Trans. Roy. Soc.*, A, **230,** 149.
Holmes, J. G., 1941. *Trans. Illum. Engng. Soc.*, **6,** 71.
Judd, D. B., 1930. *Bur. Stand. J. Res.*, **4,** 515.
Judd, D. B., 1931. *Bur. Stand. J. Res.*, **7,** 827.
McNicholas, H. J., 1936. *Bur. Stand. J. Res.*, **17,** 955.
Stiles, W. S., Bennett, M. G. and Green, H. N., 1937. *A.R.C. Technical Report R. & M.* No. 1793.
Wratten Light Filters, 1938. Eastman Kodak Co.

THE MEASUREMENT OF THE CHROMATIC AND ACHROMATIC THRESHOLDS OF COLOURED POINT SOURCES AGAINST A WHITE BACKGROUND

By N. E. G. HILL,
Royal Aircraft Establishment, Farnborough, Hants

MS. received 24 *September* 1946

ABSTRACT. Measurements were made during 1939–40 to determine the effect of background brightness on the recognition of aviation light signals. White, yellow, red, and green point-source singals were observed by monocular foveal vision against a white background whose brightness was varied from 10^{-3} to $2{\cdot}6\times10^{1}$ candles/sq. ft., a range of brightness from less than that of a starlit sky to that of a clear noon sky 20° from the sun. From the results of repeated observations of these signals curves were drawn showing the chromatic and achromatic thresholds and also the photochromatic ratio of the four colours as functions of background brightness. The curves were drawn for 50% recognition, and it is estimated that the thresholds for reasonable certainty of recognition are from three to five times those given. It is concluded that yellow is a comparatively unsatisfactory colour at both very low and very high background brightnesses.

§1. INTRODUCTION

It has been found that, when observing coloured light signals which are near the limit of visibility, the minimum signal intensity at which it is possible to recognize the colour of the signal is, in general, higher than the minimum intensity at which it is possible to detect the presence of the signal. That is to say, if the intensity of a signal be progressively reduced, the colour of the signal will disappear before the signal is lost to view. The intensities at which the colour of a signal ceases to be recognizable, and at which the signal ceases to be visible, are known as the chromatic and achromatic thresholds respectively, and the ratio of these intensities is called the photochromatic ratio of the signal. The threshold values and the photochromatic ratio are functions of the brightness of the background against which the signal is observed.

The threshold intensities of light signals are not sharply defined values, below which the signal is never seen and above which it is always seen. There is a range of intensities over which the signal will sometimes be recognized, sometimes be seen but not recognized, and sometimes not be seen at all. There are thus several ways of defining the threshold values, and we might, for instance, define the achromatic threshold either as the intensity below which the signal will never be seen, or as the intensity above which the signal will always be seen. Unfortunately these definitions, admirable in theory, do not lead to specific values in practice, and it is therefore more convenient to define the achromatic threshold at a particular background brightness as the intensity which will make the signal visible on an average of 50% of the occasions on which observation is attempted. Similarly, we shall take the chromatic threshold as the intensity at which the colour of the signal will be correctly recognized on an average of 50% of the occasions on which observation is attempted.

Some data are already available on chromatic and achromatic thresholds of monochromatic visible radiations against a black background, but only for test lights of appreciable angular size. These data have been summarized by Stiles, Bennett and Green (1937). In the absence of any data for point-source signals, tests were made at the Royal Aircraft Establishment during 1939–40 to determine the thresholds of aviation light signals. The results of these tests, which could not previously be published owing to wartime restrictions, are given in the present paper.

When dealing with point-source signals it is convenient to refer to the illumination at the observer's eye rather than to the intensity of the signal. All thresholds are therefore given as values of eye illumination in mile-candles.

§2. METHOD OF TEST

The determination of the threshold values was made by repeated observations of a series of signals, in a manner similar to that used by the author to measure the colour-recognition values of coloured light signals (1947).

The method was to fix the background at a particular brightness and then to present a succession of signals in random sequence to an observer who was required to place each signal in one of a number of colour categories or, if he failed to observe the signal, in the category "nil". The signals were of four colours, white, yellow, red, and green, and there were signals of several values of eye-illumination for each colour. The signals were repeated many times and observed by a number of different observers. The average recognitions of the various colours and of "nil" were plotted, and from the curves the chromatic and achromatic thresholds for the particular background brightness were obtained.

The tests were repeated at various values of background brightness from complete darkness to a value of brightness equivalent to that of the clear noon sky about 20° from the sun. Care was taken to maintain the background at the same white colour throughout the series of tests.

§3. DESCRIPTION OF APPARATUS

It was necessary to arrange for point-source signals to be observed in the centre of a bright background, and to ensure that the colour and intensity of the point source could be changed rapidly and accurately so that a regular succession of signals could be seen by the observer. A smooth presentation of the signals greatly relieves the observational strain in this type of test and thus leads to more reliable results.

(a) *Optical arrangement*

The general arrangement is shown diagrammatically in figure 1. The "point source" consisted of a 24-volt, 36-watt filament lamp, with a compact coiled-coil filament, mounted in a matt black screening box. Two such point sources were fitted, one for direct vision and the other to be seen by reflection at 45° in a clear glass optical flat; a ten-to-one ratio of intensity was thus obtained. A further range of intensities was obtained by means of a stepped variable sector disc giving eight values of transmission from 0·5 to 100%. Either of the filament lamps could be exposed to view by means of solenoid-operated shutters, and three coloured filters were provided so that white, yellow, red or green signals could be produced. The filament lamps were selected so that the maximum dimension of the light source, including bulb reflection, was not more than 0·1 inch.

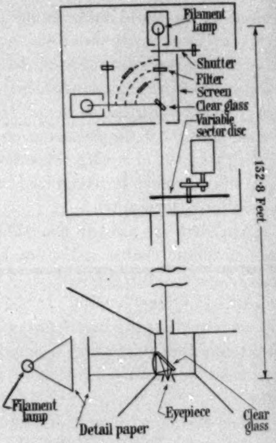


Figure 1. Arrangement of apparatus for threshold measurements.

The point-source signal was viewed by monocular vision by an observer seated 132·8 feet away. The point source therefore subtended an angle not greater than 0·25 minutes of arc. The observer's eye was located by means of a rubber eyepiece, but was not restricted by an artificial pupil. The background, which was superimposed on the signal by reflection from a clear glass optical flat, consisted of two sheets of detail paper spaced about 2 inches apart and illuminated from behind by a filament lamp. The background was viewed through a condenser lens placed so that the observer's eye was at the focus of the lens, and the observer saw an image of the bright detail paper at infinity. With this arrangement there was no difficulty in focusing the eye on the point-source signal. The bright field, which was bounded by the periphery of the lens, was circular and subtended 45° at the observer's eye.

As a help to the observer, and in order to stabilize the test results, four fixation points (not shown in figure 1) were placed at the corners of a square, of side subtending 1°·5 at the eye, and the signal appeared at the centre of the square. The fixation points were just bright enough to be seen with certainty above the background brightness. The whole apparatus was carefully screened to prevent stray light disturbing the observations.

(b) *Operational arrangements*

In order to secure rapid setting of the coloured signal, the variable sector disc was mounted on a sliding carriage whose position was varied by means of a D.C. motor controlled by a relay circuit and a set of position-selecting switches. The coloured filters were mounted in pivoted holders which could be swung in front of either of the signal filament lamps by moving three-position levers. Each lever, in addition to placing the corresponding filter in position, also closed a contact in series with the appropriate shutter solenoid.

The sector-disc and filter settings having been made, the signal was presented to the observer through a timing circuit, controlled by an electrically maintained pendulum of 1-second period. The pendulum contacts operated on a 3-second cycle in such a way that, when the push button had been pressed, a single-stroke gong warned the observer, the shutter opened for about 2½ seconds, then closed, and finally the relays were returned to rest ready for the next signal. It was found that a new signal setting could be prepared within 3 seconds, so that a continuous series of signals at 6-second intervals could be presented to the observer.

The tests were carried out with the arrangements described above at a series of background brightnesses obtained by using various sizes of filament lamp at various distances from the detail paper. In certain cases minor modifications were necessary. At the highest brightness the observer was moved to a distance of 40·1 feet from the signal source in order to obtain sufficient eye illumination from the signal. At this distance the source subtended 0·7 minutes of arc. The fixation points were suitably spaced to remain on a 1°·5 square. The brightest background was obtained by replacing the detail paper with ground glass on which was projected a defocused image of a projector lamp filament. A second condenser lens was used to flash the field of view. At the low background brightnesses it was necessary to reduce the signal intensity, and this was done by interposing a diffusing sphere in front of the direct-viewed filament lamp. In front of the sphere was placed a 0·1-inch diameter aperture.

§4. CALIBRATION OF APPARATUS

All filament lamps used during the tests, both for the signals and for the backgrounds, were calibrated for 2848° K. colour by matching with an N.P.L. standard colour-temperature lamp, and were operated throughout at that colour temperature.

The candle power of each signal lamp at 2848° K. colour was measured by standard visual photometry using the Lummer-Brodhun contrast head. In the case of the diffusing sphere arrangement, the candle power was measured using a 0·407-inch diameter aperture and the corresponding value with the smaller aperture calculated. The values of eye illumination of each signal were then calculated.

The background brightness was measured at each setting using a portable brightness photometer to transfer the brightness to the standard photometer bench.

The spectral transmission curves of the coloured filters were measured on a photoelectric spectrophotometer, and the curves are given in figure 2. The colour coordinates and total light transmission of the filters, in conjunction with a 2848° K. colour source, were calculated and are given in table 1.

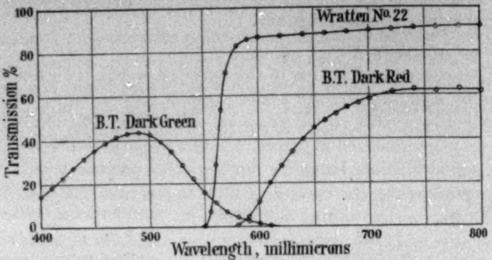


Figure 2. Spectral transmission curves of signal filters.

The white, red, and green colours are within the limits specified for aviation colours given in B.S.S. 563/1937. The yellow colour is somewhat more orange than aviation yellow.

Table 1. Colour coordinates and light transmission of coloured filters with 2848° K. colour source

Filter	Colour coordinates		Transmission (%)
	x	y	
B.T. dark green, N.P.L. 102/1924	0·184	0·392	11·90
B.T. dark red, N.P.L. 102/1924	0·693	0·307	8·18
Wratten No. 22	0·615	0·385	49·7

Colour of background is 2848° K. colour temperature.

§ 5. CONDITIONS OF TEST

The visual conditions of the tests may be summarized as follows.

Point-source signals were viewed by monocular foveal vision for about 2½ seconds against a circular white background subtending 45° at the observer's eye. The angular diameter of the point source was not more than 0·25 minutes of arc except in the case of the brightest background, when it was not more than 0·7 minutes of arc. The observer's pupil was unrestricted, but fixation points were used.

White, yellow, red, and green coloured signals of various eye illuminations were shown in succession in random order, and the background brightness was varied in steps from approximately 10^{-5} candles/ft.^{2} to 2610 candles/ft.^{2}

Observations were made by eight male observers of normal colour vision. The age groups of these observers are given in table 2.

Table 2. Age groups of observers

Age :	20–24	25–29	30–34	35–40	>40	Average : 32
Number :	1	3	2	1	1	Total : 8

§ 6. TEST PROCEDURE AND RESULTS

About five values of eye illumination were chosen for each colour, giving a total of about twenty signals for each background brightness. The settings of the apparatus for each signal were written on a small index card and the cards were shuffled to obtain a random sequence. The signals were then presented to the observer successively in the sequence given by the cards, and the observer's response to each signal was written on the corresponding card.

The observer was given a short period to become adapted to the background. In the case of the dark background, a period of 10 minutes was allowed for dark adaptation. The signals were then presented successively at 6-second intervals, and the group was repeated four times, so that about 100 signals were seen at a sitting, which occupied about 10 minutes. Short rest intervals were permitted when required by the observers. The tests were repeated at other sittings until each observer has seen each signal 25 times. The nine background brightnesses used involved a total of about 35,000 observations.

The test observations were grouped as follows :—White, yellow (including orange), red, green (including blue), nil.

The results of the eight observers were classified, added, and the percentage recognition of each of the above groups calculated. These recognition percentages are plotted as ordinates against eye illumination, E in mile-candles, as abscissae in figures 3, 4, 5, and 6 for signals which were actually white, yellow, red, and

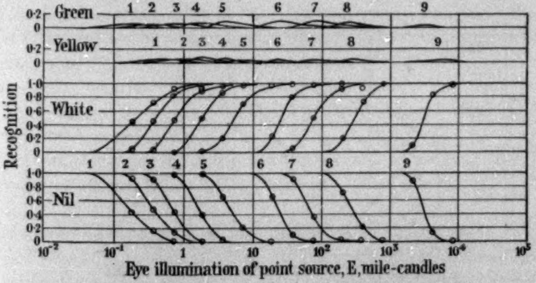


Figure 3. Recognition curves for white point source.
Curve numbers refer to background brightnesses in table 3.

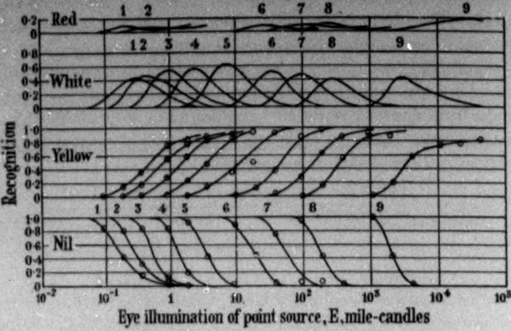


Figure 4. Recognition curves for yellow point source. Curve numbers refer to background brightnesses in table 3.

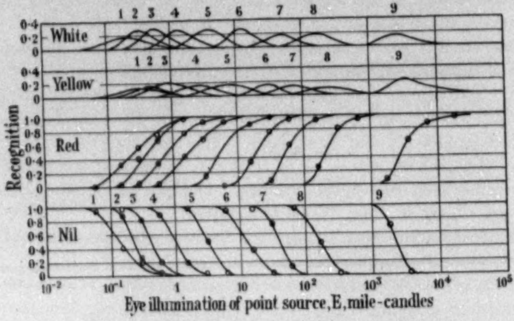


Figure 5. Recognition curves for red point source. Curve numbers refer to background brightnesses in table 3.

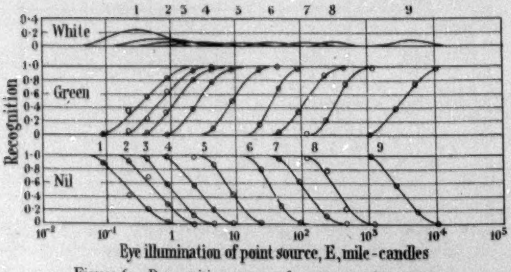


Figure 6. Recognition curves for green point source. Curve numbers refer to background brightnesses in table 3.

green respectively. The figures show families of curves, each curve at a constant background brightness, B, whose value is given in candles per square foot in table 3.

Table 3. Background brightnesses corresponding to curve numbers in figures 3–6

Curve number	Brightness, B (candles/ft^2)	$\text{Log}_{10} B$
1	Approx. 10^{-5}	$\bar{5}\cdot 0$
2	0·0111	$\bar{2}\cdot 05$
3	0·0508	$\bar{2}\cdot 71$
4	0·298	$\bar{1}\cdot 47$
5	1·75	0·24
6	10·3	1·01
7	47·1	1·67
8	292	2·47
9	2610	3·42

The achromatic threshold values for white signals can now be found by considering the "nil" recognition curves in figure 3, for evidently the 50% chance of detecting a signal is the same as the 50% chance of not seeing it. Hence the 50% "nil" ordinate gives the threshold value of eye illumination of the signal corresponding to each value of background brightness. Similarly the chromatic threshold for each value of background is obtained by reading off the values of illumination corresponding to 50% recognition of the true colour of the signal, in this case white, in figure 3.

It is clear from the general form of the "white" and the "nil" curves in figure 3 that threshold values based on the certainty of seeing or recognizing the signal, or on the certainty of not seeing or recognizing it, cannot be obtained with any reasonable precision. The reason for choosing the 50% recognition criterion for threshold values is thus apparent. It is however possible, and indeed of some interest, to find the achromatic and chromatic thresholds for 10% and 90% recognition; these values can be obtained from figure 3 fairly satisfactorily, bearing in mind that the 10% and 90% achromatic thresholds correspond to 90% and 10% recognition of "nil" respectively. In figure 7 the threshold values of illumination of white signals are plotted as functions of background brightness for 10%, 50% and 90% recognition.

In a similar manner, the thresholds for yellow, red, and green point sources are shown in figures 8, 9 and 10, the values being obtained from figures 4, 5 and 6 respectively.

The relation between the achromatic thresholds of the four signal colours is shown in figure 11, and between the chromatic thresholds in figure 12, for 50% recognition.

The photochromatic ratio, p, was calculated by taking the ratio of the 50% chromatic to the 50% achromatic threshold for each colour, and values of $\log p$ are plotted against values of $\log B$ in figure 13. In effect, figure 13 represents the ratio of figures 11 and 12.

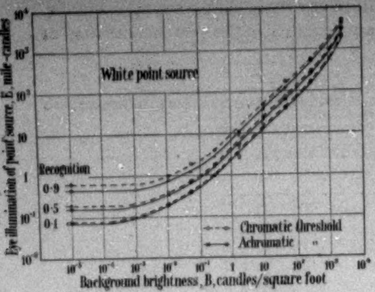


Figure 7. Threshold values of white point source.

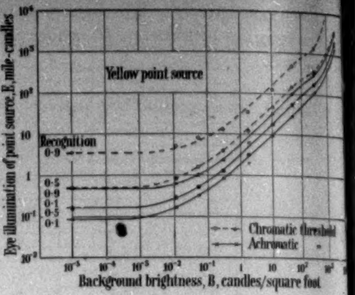


Figure 8. Threshold values of yellow point source.

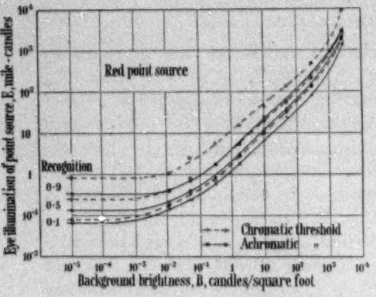


Figure 9. Threshold values of red point source.

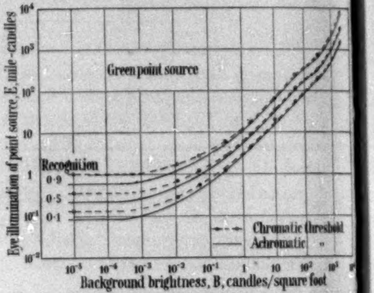


Figure 10. Threshold values of green point source.

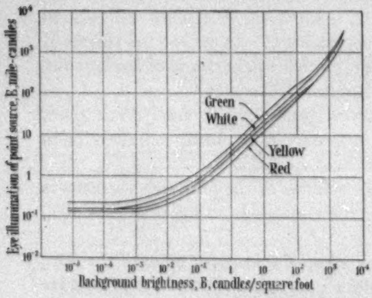


Figure 11. Achromatic threshold values for 0·5 recognition.

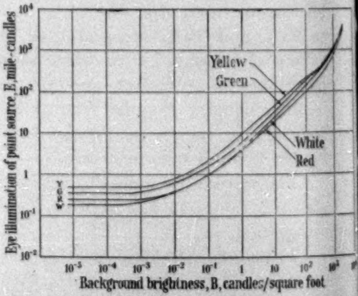


Figure 12. Chromatic threshold values for 0·5 recognition.

§7. DISCUSSION OF RESULTS

The achromatic and chromatic threshold curves given in figures 11 and 12 enable a prediction to be made as to whether a signal of a particular colour and eye illumination will be visible against a given background. Alternatively, for given background conditions, it is possible to estimate the illumination required to make a signal visible or recognizable, and hence the intensity of signal required to cover a particular range. It must be remembered, however, that the curves in figures 11 and 12 are drawn for threshold values at which there is an even chance that either the signal will or will not be observed in the case of the achromatic threshold, or that the signal will or will not be recognized correctly for colour in the case of the chromatic threshold. It will be seen from figures 7–10 that there is a considerable range of uncertainty both for achromatic and chromatic recognition, and that the uncertainty range is in general greater at the very low backgrounds. The 90% achromatic threshold varies from about 3 to about 1·5 times the 50% threshold with increasing background brightness, and the 90% chromatic threshold from about 3 times to about twice the 50% threshold except in the case of yellow, whose 90% threshold is much higher. It thus appears that, for certainty of observation, a signal would need to be less above the threshold at high background brightnesses than at low ones. This, however, is not the case in practice, because the data given here were obtained under observational conditions which did not require the observer to search his field of view for the signal, whereas, under normal conditions in aviation, the observer does not know precisely where to look. Furthermore the eye is assisted in its search at low background brightness by the extra foveal sensitivity of the retina when the eye is dark adapted (i.e. the background less than 10^{-3} candles/sq. ft.). Thus, under practical conditions, the uncertainty thresholds are likely to be from 3 to 5 times the thresholds given in figures 11 and 12.

The values of recognition for each colour group form families of related curves in figures 3–6, and the experimental points fall very well on to the individual curves. There is, therefore, an indication that the results are self-consistent, and also that a sufficiently large number of observations was made to yield statistically satisfactory averages.

The curves show the extent to which colour confusion occurs when the illumination of the signal is below the certainty level. Thus in figure 3 the white point source receives a certain amount of green and yellow recognition, but never more than 10% of either colour. The red point source in figure 5 sometimes receives as much as 30% white recognition and sometimes as much as 20% yellow recognition, although the two colour confusions do not occur together. The green point source in figure 6 may have nearly 35% white recognition against the dark background, but this confusion falls to less than 10% as the background brightness is raised. There is no indication of any confusion whatever between green and red for either the red or the green signals.

The curves in figure 4, for the yellow point source, exhibit rather different characteristics from those for the other three coloured signals. When the signal illumination is below the chromatic threshold, the recognition of the signal as white may be more than 60% at the medium background brightnesses, a value very much greater than for any of the other there signal colours. The red recognition reaches values of 10%, much the same as in the case of white signal, but, unlike that case, the red recognition curve, having reached a maximum, falls and then rises again to a second maximum at an illumination corresponding to the highest yellow recognition.

The important feature of this particular yellow signal (Wratten No. 22) is therefore that, at both low and high background brightness, its recognition as yellow fails to reach 100%, even when the illumination is well above the chromatic threshold, because of confusion with red. The author's previous work on colour recognition (1946) showed that yellow is a comparatively unsatisfactory signal colour for point sources against a dark background, and that the particular yellow now under discussion was likely to be confused with red. The present results confirm this view, and also show that this yellow is equally unsatisfactory against very bright backgrounds.

Figures 11 and 12 reveal a curious bend in the curves for the green and yellow signals at a background brightness of about 500 candles/square foot; there is no trace of any similar effect with the white or red signals. The results of some threshold measurements on a green signal made some years before the present tests, using an extinction method, suggest that the valve of background brightness at which the bend occurs is a characteristic of the individual observer. It therefore seems likely that the occurrence of the bends in the curves for the green and the yellow signals is caused by certain properties of the retina. A very interesting theory of rod-and-cone sensitivity has been put forward by Stiles (1939) which may provide an explanation, but, owing to the complexity of the theory, it has not so far been possible to apply it to the present data.

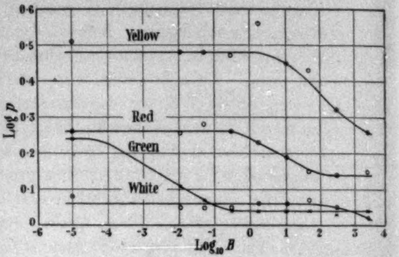


Figure 13. Photochromatic ratio for 0·5 recognition.

The photochromatic ratio curves for 50% recognition, shown in figure 13, exhibit certain interesting features. In spite of the fact that the scale for log p is rather extended, the points are found to lie very closely on smooth curves. The ratio, p, for a white signal is little greater than unity, and the ratio for green is also near unity except for very dark backgrounds, when the ratio rises to about 1·7. The value of p for red is about 1·8 for low backgrounds and falls to 1·4 for high backgrounds; this is not entirely in agreement with the common experience that a red signal looks red to extinction, but it is probably due to red-yellow confusion lowering the red recognition. The value of p for yellow is about 3 at low backgrounds and continues level until at bright backgrounds the ratio falls sharply. The existing data on the photochromatic ratio are scanty, but such data as exist do not appear to contradict the results embodied in figure 13.

Certain data exist for achromatic thresholds for dark-adapted foveal vision, and the mean values derived from these data by Stiles, Bennett and Green are: white, 0·24; red, 0·14; green, 0·32 mile-candles. Referring to figure 11, the achromatic thresholds for dark backgrounds are: white and yellow, 0·16; red, 0·14;

green, 0·22 mile-candles. Thus the present tests give the same value for red, but lower values for white and green thresholds, although the white and green thresholds are in the same ratio in each case. It is possible that a certain amount of extra-foveal recognition has occurred, lowering the white and green values but leaving the red unaffected.

§ 8. ACKNOWLEDGMENT

The author wishes to thank the Chief Scientist, Ministry of Supply, for permission to publish this paper.

REFERENCES

Hill, N. E. G., 1947. *Proc. Phys. Soc.*, **59**, 560.

Stiles, W. S., 1939. *Proc. Roy. Soc.*, B, **127**, 64.

Stiles, W. S., Bennett, M. G. and Green, H. N., 1937. *A.R.C. Technical Report R. & M.* No. 1793.

A TIME MICROMETER OF HIGH ACCURACY

By E. A. NEUMANN,

Scophony Research Laboratories, Wells, Somerset

MS. received 12 *November* 1946

ABSTRACT. A water-ethyl alcohol mixture having a zero temperature coefficient of ultrasonic velocity over a certain temperature range having been discovered, the development of an accurate time micrometer using such a mixture was attempted but was found to meet with difficulties due to partial evaporation tending to alter the composition of the liquid. Further research, however, led to the discovery that at an elevated yet convenient temperature—of 72°·7 c.—water itself displays a zero temperature coefficient of ultrasonic velocity over a useful range. A time micrometer using distilled water was therefore developed.

§ 1. INTRODUCTION

Scophony Ltd. in pre-war days developed their television receiving system which was based on the Debye-Sears effect of light diffraction by ultrasonic waves in a liquid (Scophony Ltd. and Jeffree, 1934). Experiments were carried out in this connexion for the purpose of discovering a liquid in which the speed of the ultrasonic waves would be constant over a reasonable range of temperatures. It had been found that the temperature coefficient of ultrasonic velocity in water displayed an anomalous behaviour; whereas in other pure liquids the velocity fell with rising temperature, in the case of water it increased. Efforts were therefore made to find a mixture of water and some suitable liquid in which the temperature coefficients would just compensate each other to give a zero temperature coefficient. Ethyl alcohol, well known to be miscible in all proportions with water, was tried and found suitable (Scophony Ltd. and Jerram, 1940), a mixture of ethyl alcohol and distilled water containing 16% of alcohol having a zero temperature coefficient of ultrasonic velocity at temperatures at and near 20° c.

It was quickly realized that the possibility of controlling the temperature dependence of ultrasonic velocity opened up a wider field of applications for ultrasonic waves than the one for which this possibility had originally been sought. Thus, Scophony Ltd. and Dodington (1940) suggested the use of an ultrasonic cell as a frequency stabilizer in an oscillator circuit. Again, on the suggestion of A. F. H. Thomson, formerly of the Scophony research staff, the Ministry of Supply approached the company with the suggestion that they should develop an instrument for the very accurate measurement of short time intervals, using ultrasonic waves travelling over a variable and accurately measurable distance at a known velocity kept constant within exceedingly close limits. The accuracy required was, in the course of the work, specified as *ca.* $\frac{1}{20}$ microsecond at any part of the scale, which was to be calibrated from 5 to 240 microseconds.

§2. GENERAL DESCRIPTION

The instrument which was eventually developed, and which was of the same general type as that initially envisaged, is illustrated in figure 1. Here 1 is a piezo-electric crystal having electrodes in the form of metal coatings and which, when driven by a pulse from a suitable oscillator, will vibrate and thereby generate a train of waves in the liquid 2. This train of waves, after travelling towards a plane reflector constituted by the surface of steel plunger 3, is reflected by it and returned to the crystal 1 where it produces a second pulse. The two pulses are suitably amplified and made visible on the screen of a cathode-ray tube; plunger 3 is moved by means of micrometer screw 4 rotated via gear wheels 8 and 9 by a manually operated shaft 7, until the distance between the two deflections on the cathode-ray tube screen due to the two pulses is the same as the distance corresponding to the time interval which it is desired to measure. This time interval will thus be related to a length on the micrometer scale (which, in the device

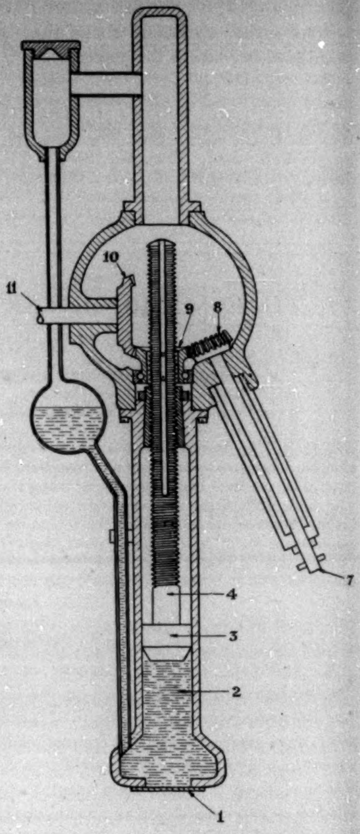


Figure 1. Sectional sketch of time micrometer.

shown, is connected to the micrometer screw 4 by means of another gear wheel 10 engaging gear wheel 9 and operating the spindle 11 of a suitable indicating device not shown). It will be seen that this method of measurement partakes of the advantages of a compensation method (such as a measurement on a Wheatstone bridge) in that any non-linearities in the time base of the cathode-ray tube are ineffective, the time measurement being reduced to the *equalization* of, instead of ordinary comparison between, two distances on the cathode-ray tube screen.

The remaining parts of the instrument illustrated in figure 1 are self-explanatory.

§ 3. DEVELOPMENT AND TESTS OF THE INSTRUMENT

The chief task in developing the instrument consisted (*a*) in a thorough investigation into the way in which the accuracy of measurements is affected by temperature fluctuations and into means to overcome the difficulty so caused, and (*b*) in a sufficiently precise determination of the ultrasonic velocity under operating conditions.

For this purpose, measuring apparatus was developed comprising two piezo-electric crystals of a standardized frequency of 18 Mc./sec., one of which was fixed near one end of a trough in which ultrasonic waves were to be produced, while the other crystal was mounted on a carriage movable along through the liquid contained in the trough. Both crystals were provided with metal coatings acting as electrodes, the first one acting as a transmitter of ultrasonic waves, for which the second acted as a receiver. The transmitting crystal was driven from a 100 Kc./sec. temperature-controlled quartz bar oscillator, the ouput of which underwent a number of stages of frequency multiplication to arrive at the required 18 Mc./sec. The output of the oscillator, in addition, underwent frequency division down to 50 c./sec., which frequency was used to drive a domestic clock, and by comparing the readings of this with the Greenwich time signals, the crystal driving frequency could be checked with abundant accuracy. The output of the receiving crystal was passed through an amplifier specially designed to ensure that its output voltage amplitude was constant and that the phase of this voltage remained fixed relative to the phase of its input. This output voltage, together with a portion of the voltage driving the transmitting crystal, was applied to a valve phase comparator, the output of which fed a meter and counter. If the movable carrier with the receiving crystal was moved towards the transmitting crystal, the needle of the meter fluctuated over almost the whole scale as the phase of the waves at the face of the receiving crystal relative to that of the waves leaving the surface of the transmitting crystal changed through 2π. The counter was arranged to increase its reading by one unit per 2π period. The receiving crystal carriage was fitted with a stop consisting of an insulated micrometer head. An 18-inch length standard bar calibrated by the National Physical Laboratory was used, its face nearest the transmitting crystal resting against a stop provided at that end. The receiver crystal carriage was driven automatically towards the other end of the standard bar, the counter operating each time the receiving crystal face advanced by one whole wave-length, until electrical contact with the standard bar was established, when the carriage was immediately brought to rest by the driving clutch being disengaged. The micrometer stop was then adjusted, moving the carriage a very short distance further towards the standard bar until the next operation of the counter occurred. The counter and micrometer were then read, after which the standard bar was swung clear of the carriage, which thereupon proceeded to advance at a speed of $\frac{1}{2}$ mm. per second, until it touched the stop near the transmitting crystal end, against which the standard bar had previously rested. The micrometer was then adjusted again until the counter operated, whereby it was ensured that a whole number of wave-lengths had been traversed in the run of the carriage. This number was obtained by subtracting the first from the second counter reading. The corresponding distance traversed was obtained as the sum of 18 inches and the difference in the micrometer readings, the measurement thus supplying all the data required for determining the ultrasonic velocity; if the distance traversed is l, the number of waves in it N, and the frequency f, then the veolcity, ${}_TV^c$, with indices T and c indicating its dependence on temperature and alcohol concentration, is equal to ${}_TV^c = l/N \, . \, f$.

In the experiments carried out, l, N and f were all determined to an accuracy better than 1 part in 10,000. The temperature of the liquid was kept constant by circulating thermostatically controlled water through the double walls of the trough provided for this purpose. Several precision thermometers were immersed in the liquid, and frequent checks of the alcohol concentration were carried out gravimetrically. This latter point, however, proved one of the main difficulties attending both the preliminary experiments and the proposed design of the actual instrument, as ethyl alcohol, as is well known, evaporates at a considerably higher speed than water, and the strength of the mixture was thus strongly inclined to alter. Thus it was found that at *ca.* 25° c. the ultrasonic velocity in a mixture containing approximately 16% of alcohol changed by *ca.* $2\frac{1}{2}$% in 18 hours.

In order, therefore, to arrive at a reliable figure for the ultrasonic velocity, several series of experiments were required, and were carried out, as follows:—

I. So as to arrive at a correction for partial evaporation, the trough was closed and sealed, a mixture comparatively rich in alcohol (over 20%) was placed in the trough and its concentration gradually and continuously reduced by the addition of water. Samples of the mixture were abstracted from the trough at intervals and its constitution determined gravimetrically. The changes in number (and fractions) of waves were continuously read from the phase comparator.

II. The actual velocity measurement had to be carried out with the trough open, as it was not otherwise possible to move the carrier. The concentration of the water-alcohol mixture was gravimetrically determined at the instants beginning and ending each run, and a correction derived from a curve representing the results of the measurements carried out under I was applied to each number of waves, as follows:—If T and c are, as before, temperature and alcohol concentration of the mixture, l is the accurately known length of approximately 18 inches as explained before, L is the distance between the crystals at their near position, ${}_TN^c_L$, ${}_TN^c_{L+l}$ etc. are the numbers of wave-lengths at temperatures, concentrations and distances indicated by the several indices, and ${}_TN_m$ the measured number of wave-lengths, then

$$\begin{aligned} {}_TN_m &= {}_TN^c_{L+l} - {}_TN^{c-\Delta c}_L \\ &= {}_TN^c_{L+l} - {}_TN^c_L + \int_c^{c-\Delta c} \frac{d({}_TN^c_L)}{dc}\, dc, \end{aligned}$$

whence

$$ {}_{T}N^{c}_{L+l} - {}_{T}N^{c}_{L} = {}_{T}N^{c}_{L} = {}_{T}N_{m} - \int_{c}^{c-\Delta c} \frac{d({}_{T}N^{c}_{L})}{dc}\, dc. $$

Measurements were carried out at an approximately constant temperature of 25° c., a small correction (of 0·0016 inch) being applied to the length of the standard bar, which had been calibrated at 62° F. = 16°·5 c. or 8°·5 below the measuring temperature. The final result of this series of measurements, with corrections applied, is shown in figure 2, which shows that the required accuracy was obtained, deviations from the mean straight line not exceeding 1 part in 10,000 represented by 0·16 metres/sec. in the velocity ordinate.

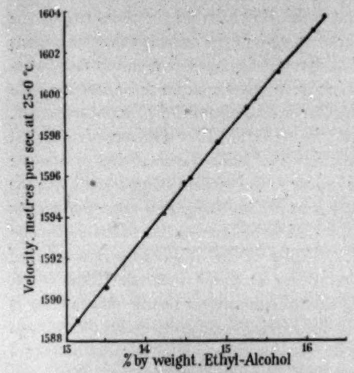


Figure 2. Velocity of ultrasonic waves in water-ethyl alcohol mixture, as dependent on concentration.

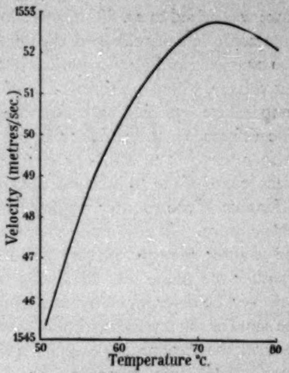


Figure 3. Dependence on temperature of the ultrasonic velocity in distilled water.

III. Measurements at different temperatures and different concentrations were also carried out with the trough closed and sealed. From these measurements a few facts emerged at once: (*a*) It was found that the attenuation of supersonic waves was very rapid below 20° c. (*b*) Different mixtures have a zero velocity-temperature coefficient at different temperatures. (*c*) No mixture was found for which the region of substantially zero velocity-temperature coefficient covered an extended temperature interval; the useful interval in fact proved to be substantially constant over the whole range investigated and to amount to 6° c. if a variation of not more than 1 in 10,000 was permitted.

Because of (*a*) and (*c*), it was decided to embody some temperature control in the final instrument to keep the temperature at a somewhat elevated point. There remained the difficulty of the mixture being inclined to change its composition, due to the different rates of evaporation of its components.

To overcome this difficulty, it was suggested either to use a mixture of which only the water component was liable to evaporate to any sensible degree, e.g. sodium iodide/water or glycerine/water, and periodically to "top up" the mixture with water, or to use a mixture both components of which evaporate at or near the same rate, e.g. to mix water with propyl alcohol (the vapour-pressure curve of which is almost identical with that of water). The latter method was favoured.

There followed, however, a discovery which, besides being interesting in itself, considerably simplified the problem. This was that pure water behaved in a way similar to that in which the water-ethyl alcohol mixtures had been shown to behave: the ultrasonic velocity in it as plotted against its temperature went through a maximum.

The following method was now adopted for arriving at a curve accurately relating ultrasonic velocities to temperatures. The receiving crystal was placed about 25 inches from the transmitting crystal and fixed, and the trough, filled with distilled water of 0·6 megohms per cm^3 at 20° c., was covered. The temperature was raised to *ca.* 80° c. and allowed to fall slowly. As, in consequence, the supersonic velocity changed, the number of wave-lengths between the two crystals also changed, leading to a gradual relative change of phase between input and output voltage, whole multiples of 2π of which were registered on the counter. The temperature was read each time the counter operated until 51° c. was reached. The difference in the counter reading, ΔR_1, from that at 51° c. was recorded against temperature. The whole procedure was repeated with the receiving crystal the exact length of the 18-inch standard bar nearer to the transmitting crystal, and the difference of the counter reading from that at 50° c., ΔR_2, also recorded against temperature. In evaluating the results, it had to be borne in mind that, although the counter recorded phase change in multiples of 2π, it gave no indication of whether the phase was advancing or retarding at these instants. This could, however, be determined by ascertaining whether the meter needle was moving in the same or the reverse direction, as when the carriage was given a slight movement towards the transmitting crystal; it was in this way known whether the number of wave-lengths was increasing or decreasing at any temperature. With this knowledge it was possible to ascertain ΔN_1 and ΔN_2, the difference in the number of wave-lengths in the distance separating the crystals at the temperature T, and at 51° c. and 50° c. respectively, for the two separations 25 inches and 7 inches approximately. It was found from curves relating ΔN_1 or ΔN_2 to T that the maximum of ultrasonic velocity in distilled water of the stated degree of purity occurs at about 72°·7 c. At this temperature a complete run over substantially the length of the standard bar was taken, as explained earlier in this article, to arrive at the ultrasonic velocity at this particular temperature, which was found to be equal to 1552·7 metres per second. Velocities at other temperatures were now derived from the measurement leading to the value at 72°·7 c. and from the curves relating ΔN_1 and ΔN_2 to T, thus saving a considerable amount of time as compared with that which would have been needed to measure the velocities at various temperatures in the same way as at 72°·7 c.

Let x be the exact length of the 18-inch bar, and the several symbols and indices having the meanings explained hereinbefore, then

$$\Delta N_1 = {}_{51}N_{L+x} - {}_{T}N_{L+x},$$

$$\Delta N_2 = {}_{50}N_L - {}_{T}N_L,$$

and

$${}_{T}N_x = {}_{T}N_{L+x} - {}_{T}N_L = ({}_{51}N_{L+x} - {}_{50}N_L) - (\Delta N_1 - \Delta N_2).$$

$_{72\cdot7}N_x$ was measured when the ultrasonic velocity at 72°·7 c. was determined, and was found to amount to 5300·4 wave-lengths, and, at the same temperature, $\Delta N_1 - \Delta N_2$ was found to amount to 25·6. It follows from this that

$$_TN_x = 5326{\cdot}0 - (\Delta N_1 - \Delta N_2),$$

and, therefore,

$$V_T = \frac{fx}{5326{\cdot}0 - (\Delta N_1 - \Delta N_2)} \text{ inches per second}$$

$$= \frac{8229{\cdot}6 \times 10^3}{5326{\cdot}0 - (\Delta N_1 - \Delta N_2)} \text{ metres per second.}$$

V_T is shown plotted against T in figure 3 (Jones and Gale, 1946).

After this discovery had been made, it was decided to operate the time micrometer with a distilled-water filling and at *ca.* 73° c., and the instrument was constructed accordingly. The thermo-controls of the micrometer were so devised that on starting operations a primary or auxiliary heater was put into action which quickly raised the temperature of the whole to a temperature in the vicinity of the correct operating value of 73° c.; after 30 minutes, the supply to the auxiliary heater was automatically switched off by one of the bimetallic switches incorporated in the device, the temperature being subsequently maintained solely by the maintenance heater. The correct operating temperature of 73° c. was attained after approximately another ten to fifteen minutes.

The traversal of the plunger (3 in figure 1) was operated by gearing driven by a hand-wheel; gear ratios 1 : 1 and 4 : 1 could be obtained by pulling out and pushing in the hand-wheel. Provision was made for automatically disengaging the drive from the plunger at the two ends of the range over which it was to operate, and for re-engaging it on reversal of the hand-wheel.

The time range to which the range of plunger travel was to correspond had been specified by the users as being from 5 to 240 microseconds. Keeping in mind that the distance from crystal to plunger surface was travelled over twice by the ultrasonic waves (once before and once after reflection), this corresponds to a distance from crystal to plunger of from approximately 4 mm. to approximately 192 mm.

The counter was so calibrated that one unit on it corresponded to about 1·2 microseconds (the exact figure had been specified by the users), and that $\frac{1}{20}$ of a unit could be read on any part of the scale. By the side of the counter was a thermometer, underneath a water-level indicator, near the top of the front panel of the apparatus a green "tell-tale" lamp indicating correct operating conditions, and near the lower edge terminals and power switch. A push-button served as an automatic cut-out and reset switch which operated the primary heater.

The internal mechanism of the micrometer was stainless steel throughout to prevent corrosion.

ACKNOWLEDGMENTS

Research preparatory to the design of the instrument was carried out by A. J. Gale, to whom several suggestions, including that to use pure water, are due, P. L. F. Jones being in general charge of the research and development. The

author, in writing the present article, has freely drawn on a report by the latter. The details of the actual design were dealt with by A. E. Adams.

Thanks are due to the Director of Scientific Research, Ministry of Supply, for permission to publish the results of this work.

REFERENCES

Jones, P. L. F. and Gale, A. J., 1946. *Nature, Lond.*, **157**, 341.
Scophony Ltd. and Dodington, S. H. M., 1940. Brit. Pat. No. 573,269.
Scophony Ltd. and Gale, A. J., 1947. Brit. Pat. No. 582,435.
Scophony Ltd. and Jeffree, J. H., 1934. Brit. Pat. No. 439,236.
Scophony Ltd. and Jerram, C. F., 1940. Brit. Pat. No. 534,448.
Scophony Ltd. and Thomson, A. F. H., 1947. Brit. Pat. No. 582,434.

COLORIMETRY IN THE GLASS INDUSTRY

By J. G. HOLMES

A lecture given to the Colour Group 14 *November* 1945; *MS. received* 27 *September* 1946

ABSTRACT. Recent knowledge of the glassy state has enabled the theories of modern colour chemistry to be applied to glass, and developments in colorimetric technique have put the design and performance of coloured glasses on a quantitative basis. The methods of colour measurement particularly suited to transparent media are described, together with rapid approximate methods of calculation. The properties of the important colouring oxides are given, and the effects of concentration, thickness and illuminant are discussed. The colours and reflexion-factors of bloomed glass surfaces, the properties of some special colour filters and a basis for specification of coloured glasses are briefly described.

§ 1. INTRODUCTION

Although the glass industry is by no means a large user of colorimetric methods, this lecture must be restricted to a few of the applications of colorimetry and, therefore, to those with which the author is most familiar. Amongst the important items which are not discussed are the colours of decorative and domestic glass, coloured opal glass and similar surface colours, photoelectric colorimetry and the terminology of glass colours. The subjects discussed will include a very brief statement of the background knowledge of coloured glass, the methods of measurement and calculation appropriate to a transparent medium and one or two points of interest, including some special colour filters, the colours of "bloomed" lenses with non-reflecting films and the basis of specification for coloured light signals. This last item is the one to which colorimetric methods are most widely applied, as glasses obtained through ordinary commercial channels are not usually closely graded, and it was also one of the first industrial applications of the trichromatic system agreed in 1931 by the Commission Internationale de l'Eclairage.

§ 2. THE STRUCTURE OF GLASS

It is not enough to say that glass is a "super-cooled liquid". It has recently been defined by Scholes (1945) as "an inorganic product of fusion which has cooled to a rigid condition without crystallizing". It is typically hard and brittle, but it may be colourless or coloured, transparent or opaque. Its structure is similar to that of the liquid state characterized by so high a viscosity that it is for all practical purposes rigid (Morey, 1938). Just as liquids are analogous to crystals, so there may be a close similarity between the arrangements of atoms in a glass and in a crystal, even though there is no crystalline structure in glass.

The current theory of glass regards it as a three-dimensional network consisting mainly of silicon and oxygen atoms in random orientation, each silicon atom being bonded to four oxygen atoms and each oxygen atom to two of silicon, and, as the bond is very strong, the properties of silica glass are very stable. Other atoms, such as boron, which have glass-forming oxides, may take their place in the network, but mostly the other elements used in glass-making go into the holes in the network and will generally loosen the silicon-oxygen network and alter the physical properties. For example, the addition of sodium oxide to silica may be represented by sodium ions in holes adjacent to oxygen atoms which are bonded to only one silicon atom, causing amongst other things a lowering of the softening temperature and an increase in the thermal expansion coefficient. If boric oxide is added to the soda-silica glass, it forms part of the network, reducing the number of single-bonded oxygens and tightening up the whole system, reducing the thermal expansion coefficient. If lead oxide is introduced, the lead goes into the holes in the network and makes a heavier softer glass and, in general, elements of different atomic weights and different valencies and affinities will yield glasses of different properties. Some elements will give an unstable system of forces between the ions, and this instability is associated with selective absorption of light. The deepest absorption bands, which give the deepest colours when they occur in the visible region of the spectrum, are associated with ions of two different valencies of the same element, and, for example, manganese will normally give a rich purple colour, due to unstable balance between the oxidized and reduced states, but this colour is almost completely absent if the manganese is strongly reduced. Ferrous iron gives a strong absorption band in the infra-red and ferric iron gives a strong absorption band in the blue, but in practice it is difficult to obtain either complete reduction or complete oxidation, and glasses coloured with iron are subject to control of the ferroso-ferric balance.

The system of forces in the network will depend on both the composition of the network-formers and the modifiers or chromophore ions, and thus the absorption bands may be affected by the base glass as well as by the colouring ingredients. For example, cobalt usually gives a blue colour in soda-silica glasses, but it gives a reddish colour in borate or phosphate glasses which are highly acidic, and it gives a pink colour if it replaces the sodium in a soda-silica glass. Titanium has the property of modifying the network and loosening its structure, so that an ion which normally goes into a hole in the network may take up a silicon position in the network itself, and the colouring effect of the ion is greatly affected even though titanium produces no coloration itself. An example of this is a ceria-soda-silica glass which is colourless but which becomes a pronounced yellow when titania is introduced.

Raising the temperature of glass will loosen the structure and reduce the differences between the energy levels, so that the absorption bands tend to move towards longer wave-lengths which have a smaller quantum of energy.

This network-model of the glassy state is far from complete, but it provides a very useful basis for argument. It is developed in some detail in a monograph by Weyl (1944) now being published in the *Journal of the Society of Glass Technology*.

The difference between glass and crystals is illustrated by the absence of any sharp change in refractive index associated with an absorption band in coloured glass. There may be some relation between the rise in index and the rise in absorption towards the ultra-violet end of the transmission spectrum, where the absorption is due to the forces in the network rather than to a modifying ion, and if so, this would show the family relationship between the random network in glass and the regular network in crystals.

§ 3. THE PROPERTIES OF A TRANSPARENT COLOURED MEDIUM

Glass is an excellent example of a coloured material, because the effects of absorption under different circumstances can be calculated or estimated from comparatively simple data and there are no effects of texture or gloss. Colour is the subjective effect of selective absorption in the visible spectrum, and the colour name given to a glass is the complement of the colour which is absorbed. In fact, glass is a simple example of the subtractive process of producing colour, and it follows that subtractive instruments such as the Lovibond Tintometer are as suitable for measurement of coloured glasses as they are for coloured liquids.

Greater concentration of the colouring constituent in a glass, or greater thickness of glass, will give a lower transmission factor (the glass-maker's equivalent of lightness or brightness of surface colours) and will usually give a purer, more saturated colour. Considering a cobalt blue glass as an example, figure 1 shows the transmission-wave-length curves of six glasses containing increasing amounts of cobalt and figure 2 shows the colours of the light, from a source operated at a colour temperature of 2848° K., after transmission through each of the glasses. The strongest absorption band in figure 1 is at about 600 mμ, and so the colour of a pale cobalt glass in figure 2 is on the side of the source towards the complementary wave-length. As the amount of cobalt increases, the orange radiation near 600 mμ is almost completely absorbed and the absorption in the yellow-green region becomes more noticeable, so that the colour locus moves away from the yellow-green region. The darkest glass in figure 1 shows absorption of substantially all the orange-yellow-green radiation and the colour of the transmitted light may be represented by the centre of gravity of the wave-length bands from 400 mμ to about 500 mμ and from about 680 mμ to 750 mμ. In general, figure 2 shows how the hue and saturation of the transmitted light change as the amount of cobalt is increased. It may be noticed that although the process is subtractive, the author's thinking is in terms of additive mixing of the light which is transmitted.

Another important property of a glass is its transmission factor, which is the percentage ratio of the amount of light transmitted through the glass to the amount of light incident on it, evaluated in terms of the standard visibility function. The absorption of light necessary to produce colour means that coloured glasses

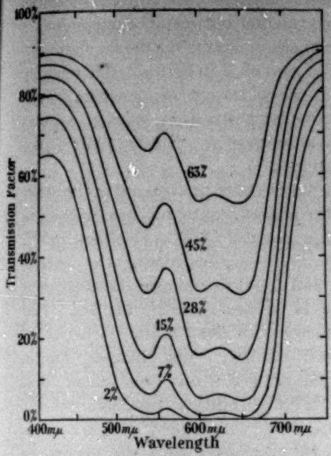


Figure 1. Transmission curves of six cobalt glasses. (The figures show percentage total transmission factor.)

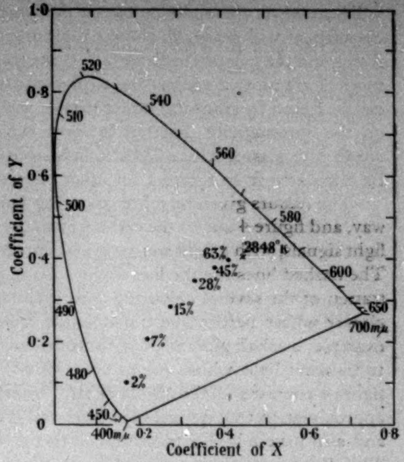


Figure 2. Colours and transmission factors of six cobalt glasses with 2848° K. light source.

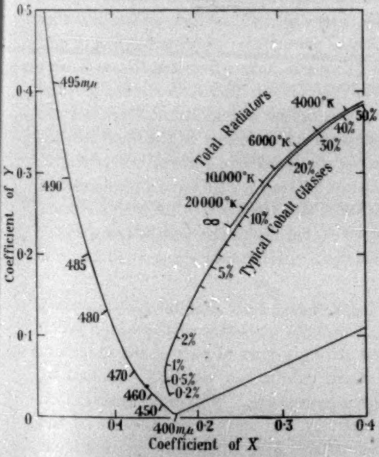


Figure 3. Colour-transmission relation for typical cobalt glass with 2848° K. light source.

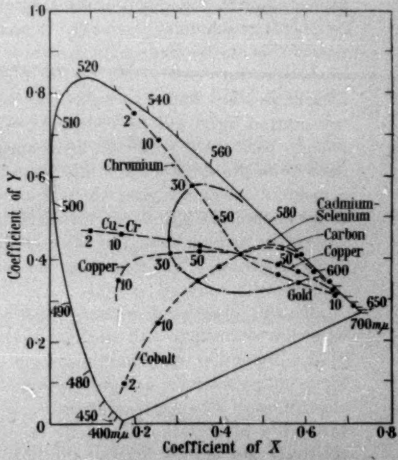


Figure 4. Colour-transmission relation for typical signalling glasses with 2848° K. light source.

will have lower transmission factors than uncoloured glasses, and a given colouring constituent will generally show a reproducible relation between the transmission factor and the colour coefficients. If the transmission factors for the six glasses in figure 1 are plotted against the coefficients of X and of Y in figure 2, this relation can be found by graphical interpolation, and a scale of transmissions can be drawn on the chromaticity diagram to show the colour-transmission locus of typical cobalt blue glasses with a 2848° K. light-source as in figure 3. Similar figures may be calculated from figure 1 for other light sources.

The colours given by other colouring constituents may be analysed in the same way, and figure 4 shows the colours given by the glasses commonly used in colour-light signals, with a light source at a colour temperature of 2848° K. (Holmes, 1937). The dashed lines are the loci of the colours given by varying thickness or concentration of the several colouring constituents, and the dots indicate the colours of glasses whose percentage transmission factor is written close by the dot. For example, a cobalt glass of 50% transmission factor with 2848° K. may be expected to transmit light whose colour is $0{\cdot}38X+0{\cdot}40Y+0{\cdot}22Z$. The chain-dot line in figure 4 connects all the points of 30% transmission, and it may be taken as a first approximation that no ordinary glass can give a colour outside this chain-dot line and also have a higher transmission factor than 30%, with a 2848° K. light source. This "maximum-transmission locus" may be compared with the maximum pigment colours calculated by MacAdam (1935), and it will be found that the yellow-orange and pure red glasses are not far removed from the theoretical maximum transmission for their colour, but the green and blue-green glasses give transmission factors much below the maximum for the same colour or, alternatively, give colours whose purity is much less than theoretically possible for the same transmission factor. It is of interest to see that the orange and red colours given by cadmium-selenium glasses can be matched by spectral colours. The explanation of this on the diagram is that these glasses absorb the short wave-length end of the spectrum completely, whilst transmitting the long wave-lengths with very little absorption, and if all wave-lengths less than about 540 mμ are absorbed, the transmitted light will be composed of wave-lengths whose colours are on the straight part of the spectrum locus and, therefore, the colour of the mixture will itself lie on the straight part of the locus and be matched by a spectral wave-length. A cadmium-selenium glass which absorbs below 540 mμ would have a transmission factor of about 55% and a colour of about $0{\cdot}585X+0{\cdot}414Y+0{\cdot}001Z$ with 2848° K., this colour being matched by the wave-length of 592 mμ in the yellow-orange part of the spectrum. The maximum theoretical transmission factor to give this colour is about 64%.

On the other hand, glasses which give green and blue colours will generally be of low saturation with artificial light. The light transmitted through a green glass will contain wave-lengths lying on the strongly curved part of the spectrum locus and, therefore, a highly saturated green colour can only be obtained by absorption of all except a narrow band of the spectrum. The light transmitted through a blue glass will usually be of relatively low saturation because of the relatively low energy level at the blue end of the spectrum from incandescent filament lamps, from which it is only possible to achieve a highly saturated blue colour by employing a glass of very low transmission factor as indicated in figure 3.

§4. THE MEASUREMENT AND CALCULATION OF TRANSMISSION FACTOR

The measurement and calculation of transmission factor is the first part of the colorimetry of glass, being simpler than the measurement of colour and yet bearing a close relationship to colour. Incidentally, the American word for transmission factor is "transmittance", and there is a risk of confusion with our word transmittance, which has a different meaning. We say that transmission factor is the ratio of the light leaving a glass to that incident upon it and that transmittance is the value which this ratio would have if there were no reflexion of light at the two air-glass surfaces. To a first approximation, the transmittance of flat glass is 1·08 times the transmission factor.

Lambert's law is strictly obeyed by all non-fluorescent glasses:—

$$I=I_0\,.\,T=I_0\,.\,10^{-D},$$

where I is the intensity of the transmitted light, I_0 is the intensity of the incident light, T is the transmission factor, and D is the optical density.

The optical density D is the common logarithm of the reciprocal of the transmission factor T. The internal optical density d (sometimes written ID) bears the same relation to the transmittance t. Thus:

$$D=\log_{10}(1/T) \quad \text{or} \quad T=10^{-D},$$
$$ID=d=\log_{10}(1/t) \quad \text{or} \quad t=10^{-d}.$$

If r is the reflexion loss,

$$T=t\,.\,(1-r),$$
$$d=D+\log_{10}(1-r),$$
$$T=(1-r)\,.\,10^{-d}.$$

Bouguer's law of variation of transmission factor with thickness (sometimes ascribed to Lambert) is obeyed provided the quality of the light is unchanged, as in truly neutral grey glasses or in monochromatic light:

$$I_x=I_0\,.\,(1-r)\,.\,(t_1)^x=I_0\,.\,(1-r)\,.\,10^{-d_1x},$$

where I_x is the intensity after transmission through a thickness x and t_1 and d_1 are the transmittance and internal density for unit thickness.

Beer's law is not generally obeyed for variations in concentration of the colouring constituent, although a modified relation can be found as indicated below.

The thickness-conversion equations for transmission of monochromatic light through thicknesses x and y can be stated as follows:

$$t_x=(t_1)^x, \qquad (t_y)^x=(t_x)^y,$$
$$d_x=x\,.\,d_1, \qquad d_y=y\,.\,d_x/x.$$

Thus internal density is proportional to thickness for monochromatic light.

In terms of transmission factor,

$$T_y=(1-r)\,.\,[T_x/(1-r)]^{y/x}$$

or

$$\log T_y=\log(1-r)+[\log T_x-\log(1-r)]\,.\,y/x.$$

These relations are old established, but have only recently appeared in technical literature (Sharp, 1942, and McLeod, 1945). The last equation can be solved quickly by several methods, and the most accurate and quick method of converting from one thickness (x) to another thickness (y) is to use two slide rules, one set for the ratio y/x and the other set to read $\log_{10}[T/(1-r)]$. This second setting can be made by first calculating the reflexion loss from the Fresnel expression $(n-1)^2/(n+1)^2$, which leads to a value 0·92 for the factor $(1-r)$ if the refractive index n is 1·516, and then setting the linear scale, usually found on the back of the B- and C-scales of a slide rule, with its zero opposite to 92 on the D-scale. The linear scale will then be the internal density, which is proportional to thickness, and the D-scale will be the percentage transmission factor, T. A tabular form of the twin slide-rule method has been described by Gage (1937).

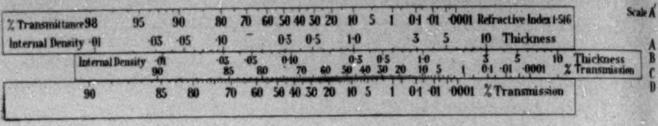

Scale A′
% Transmittance 98 95 90 80 70 60 50 40 30 20 10 5 1 0·1 ·01 ·0001 Refractive Index 1·516
Internal Density ·01 ·03 ·05 ·10 0·3 0·5 1·0 3 5 10 Thickness
A
Internal Density ·01 ·03 ·05 0·10 0·3 0·5 1·0 3 5 10 Thickness
B
90 85 80 70 60 50 40 30 20 10 5 1 0·1 ·01 ·0001 % Transmission
C
90 85 80 70 60 50 40 30 20 10 5 1 0·1 ·01 ·0001 % Transmission
D

Figure 5. Slide rule for transmission-thickness conversions.

It is not difficult to make a slide rule to solve the thickness-conversion equation directly for a given reflexion loss, and figure 5 is a sketch of such a rule which was made just before the war and which has been in almost daily use since that time. In figure 5, the slide is set for a 3 : 10 change in thickness as shown on the A- and B-scales, and the corresponding change in transmission factors may be read off the C- and D-scales, such as, for example, 50% transmission factor at 3 mm. thickness becomes 12% transmission factor at 10 mm. thickness. The A- and B-scales are actually a logarithmic ruling for the internal density, which is proportional to thickness, and the C- and D-scales are calculated from the equation. The A′-scale is the percentage transmittance corresponding to the internal density on the A-scale and the transmission factor on the D-scale. An ingenious circular form of this slide rule has recently been described by Vaughan (1944).

A very simple graphical method of calculating the effect of thickness changes is to use linear-log graph paper, suggested to the author by Dr. W. M. Hampton. In figure 6 the ordinates are the transmission factors on a logarithmic scale and the abscissae are thicknesses on a linear scale. The straight lines are the relations between thickness and transmission for two neutral glasses, both passing through 92% at zero thickness and elsewhere conforming strictly to the thickness-conversion equation above. This method has recently been published by Powell (1945). The curved line has been experimentally determined for a coloured

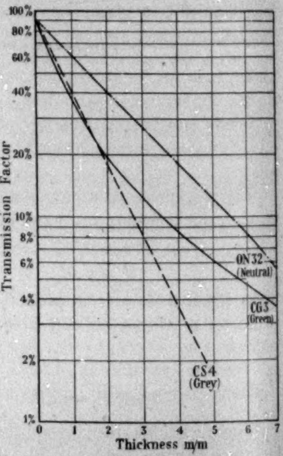


Figure 6. Graphical method for transmission-thickness conversions.

glass and it is found that all coloured glasses in which the quality or composition of the light is changing as it passes through greater thicknesses give concave curves of the type indicated. Some glasses, such as browns or blue-greens, give curves which are only slightly bent, and others, such as selenium rubies, give very strongly bent curves. The explanation of the curvature is that, as the thickness increases, the wave-lengths which are more strongly absorbed become of less importance and the light becomes relatively richer in the wave-lengths which are more freely transmitted, so that the transmittance per unit thickness is greater than it would be if the quality of the light had not changed, and so the slope of the line changes. Glasses of a given type will all show similar curves, and this method has proved most valuable in the routine control of coloured glass during the past fifteen years or so.

The departure from Beer's law may be represented in the same way, and the transmission factor for constant thickness but varying concentration may be shown as a curved line on a diagram similar to figure 6.

Emphasis has been placed on the arithmetical methods used in conversion from one thickness to another because these represent an important part of the glass-maker's colorimetric technique. The measurement of the transmission factor of glass presents no great difficulties, and it is the author's preference to employ visual methods exclusively. Photoelectric cells give quicker readings but tell lies without blushing. The reliability of flicker photometers is very dependent on the skill of the observer, and in extreme cases threefold errors have been obtained in measurement of the transmission factor of purple glasses with several observers. The most reliable method has been to use a photometer bench with a Lummer-Brodhun contrast head, two light sources of controllable colour temperature and a wide range of accurately calibrated glasses of all the common colours for use as comparison standards. The calibration is best done on a spectrophotometer, and this may of course be visual or photoelectric.

An interesting aspect of transmittance measurement is in the control of through-coloured stepped lenses (Fresnel lenses). There is a specification which defines a minimum transmittance (BSS 623–1940) and the method of measurement is to compare, on a photometer bench, the brightness of the photometer screen illuminated through the coloured glass lens and through a colourless glass plate with the brightness when the screen is illuminated through a colourless glass lens of the same pattern and a calibrated coloured glass plate. The light source is an opal lamp operated at the appropriate colour temperature and placed at the focus of the lens, conjugate to the photometer screen. Each of these assemblies is matched in turn against a coloured light on the other side of the photometer head, the properties of this light being unimportant except that there should be a reasonable colour match. The ratio of the two brightnesses is the same as the ratio of the unknown transmittance of the coloured lens to the known transmittance of the coloured glass plate, and thus a measurement which might be very complex becomes a matter of simple routine. For some types of glass the limiting colour stated in the specification can be correlated with a maximum transmittance and, as it is often found that there is a close correlation between weight or thickness and transmittance in a single batch of glasses, it is not unusual to determine the limits by careful colorimetric technique and then to carry out the routine examination of each glass with a spring balance or dial gauge. Border-line glasses would of course be subjected to colorimetric examination.

§ 5. THE MEASUREMENT AND CALCULATION OF COLOUR

The transparent nature of glass emphasizes that, in common with other coloured materials, it possesses no colour of its own, but only shows colour by its selective absorption of light passing through it. Unlike a surface colour, glass is not usually looked at, but is looked through, and it is immediately obvious that we have to measure the colour of the combination of the glass and the light source. This point is particularly important in the colorimetry of signal glasses which are used with light sources of widely varying colour-temperature and misconceptions are liable to arise.

The simplest method of checking that the colour of a glass is between defined limits is to compare the test glass with limit glasses or with a calibrated standard. The photometer bench with lamps of adjustable colour temperature and calibrated standard glasses of the same type as the test glass will give sufficient accuracy for most purposes. Where interpolation is required to a higher accuracy than by visual estimation between two colours, it is common practice to use a wedge of glass which is calibrated for colour along its length and to match the test glass against the appropriate thickness of the wedge.

Any type of colorimeter can be used for measuring the colours of glasses if there is some attachment enabling it to measure the colour of light. The Hilger-Guild instrument or the Donaldson instrument are both excellent for this purpose and the Lovibond Tintometer is very good except for the purest colours, which may require the addition of a neutral shade to bring them within the range of the instrument. The type of "colorimeter" which ought to be called an "absorptiometer" is not, of course, suitable for direct measurement of colour, but it may be used to give reliable results by the method of abridged spectrophotometry if it is suitably calibrated, preferably by reference to a known glass of the same type as the test glass.

The best method of colour measurement is to determine the transmission factor throughout the visible spectrum, because this gives the whole relevant information about the glass, and this information can be readily handled by calculation. The author's usual method is to calculate the trichromatic coefficients and the transmission factor from the spectrophotometric data for a series of thicknesses, leading to the colour-transmission relation for the glass as shown in figure 3 for cobalt glass. The only practicable experimental method for obtaining data such as this is the spectrophotometric method.

From a knowledge of the typical spectrophotometric transmission curve, the effect of change of the colour temperature of the source can be readily calculated, and curves such as those in figure 4 can be drawn for other light sources. This is particularly valuable for glasses for coloured light signals which may have any illuminant from an oil flame at 1900° K. to an arc lamp at 3500° K. and in which the effect of change of the illuminant may be greater than the whole tolerance allowed for the colour of the glass itself (Holmes, 1937). Figure 7 shows the colours and transmission factors of Aviation Green glass (Chance CG6) for three light sources, calculated from a single spectrophotometric curve by the methods described below.

In the calculation of colour it is general practice to employ the CIE trichromatic system, although a number of users of coloured glass continue to describe its colour in terms of wave-length and saturation. The author's calculations are on the weighted ordinate method rather than the selected ordinate method (Hardy, 1936) because the former gives a higher accuracy with less labour. The weighted ordinate method was described by Smith and Guild in 1931 and consists of two stages for each of the three primaries, the first stage being to multiply the transmission factor (T_λ) by the energy level of the light source (E_λ) and by the distribution coefficient ($\bar{x}_\lambda$, etc.) recommended by the CIE in 1931, making this calculation for each wave-length at intervals of, say, 0·01 micron through the spectrum, and the second stage being to add the products to give the trichromatic equation for the colour. The first stage can be greatly simplified by constructing a permanent table of products, either by calculating machine or by careful use of a slide rule, and the second stage can then be done by adding selected products on an adding machine, so avoiding the necessity for laborious multiplication. Table 1 shows some of the entries in the master table for X, Y and Z for Illuminant A. In the master table, the required product ($T \, . \, E \, . \, \bar{x}$ etc.) has been calculated for each percentage transmission from 1% to 90% and for each wave-length at intervals of 0·01 micron, based on the figures given by Smith (1934). The appropriate products are chosen from each column and added by machine to give the coefficients of the trichromatic equation. Although the products are only worked out for each 1% transmission, it is simple to make calculations for decimal percentages by moving the decimal point.

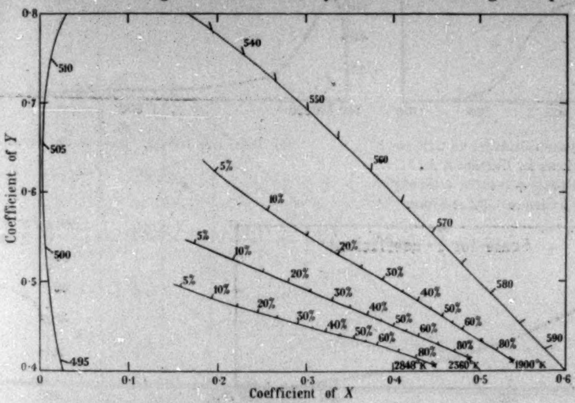
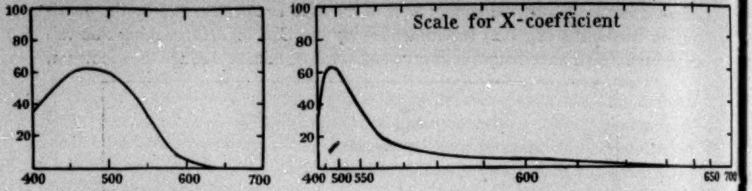


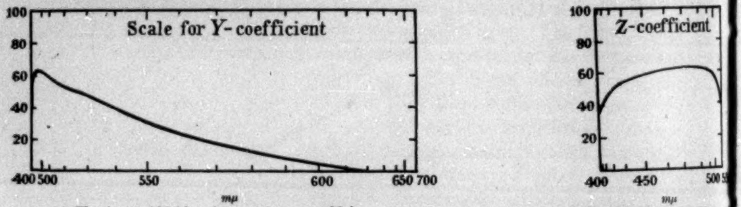
Figure 7. Colour-transmission relations for aviation green glass (Chance CG6).

The weighted ordinate method can also be applied graphically, using scales which have been published in an earlier paper (Holmes, 1935) as illustrated in figure 8. Figure 8 (*a*) shows the curve of transmission factor and wave-length for a blue-green glass, plotted with linear scales in the ordinary way. Figures 8 (*b*), 8 (*c*) and 8 (*d*) show the same curve plotted with a linear scale of transmission factors but with non-linear scales of wave-lengths. The spacing of each wave-length scale has been calculated according to the energy distribution of the light source (in this case, Illuminant A or 2848° K.) and the distribution coefficients of the standard observer for each of the three primaries. It follows that the area under each of the three curves is proportional to each of the three trichromatic coefficients of the colour and if the areas are reduced to unit sum, the result will be the unit trichromatic equation. The area under the curve for the *Y*-coefficient is, of course, the transmission factor of the glass. The advantages of the method are firstly that you can see what you are doing and secondly that a reasonably accurate result can be obtained with very few observed points on the transmission curve. The method is quick if permanent ruled charts are kept, but it is subject to all the usual errors of the measurement of area by a planimeter. Incidentally, it may be argued that this is a graphical form of the selected ordinate method, rather than the weighted ordinate method, and it may be derived from either.



(*a*) Graphical calculation of CIE coefficients for Illuminant A.
Source: $0{\cdot}448X + 0{\cdot}407Y + 0{\cdot}145Z$.
Filter: Chance OB2 at 2 mm.

(*b*) Total area 109·85. Area under curve 12·4.



(*c*) Total area 100·00. Area under curve 22·0.

(*d*) Total area 35·58. Area under curve 20·6.

Figure 8. Graphical method for colour calculations.
Unit equation $0{\cdot}225X + 0{\cdot}400Y + 0{\cdot}375Z$. Transmission factor 22·0%.

The idea of a unit equation is sometimes rather difficult at first for a student and it may be suggested that a "percentage equation" is easier to understand. In the example quoted in figure 8 the light transmitted by the glass is represented by the three areas: $x' = 12{\cdot}4$, $y' = 22{\cdot}0$, $z' = 20{\cdot}6$.

If the total is regarded as 100%, which is immediately understandable to any student, it is clear that this corresponds to 22·5% of *X*, 40% of *Y* and 37·5% of *Z*, and he may then see that this is another way of expressing the unit equation:

$$C = 0{\cdot}225X + 0{\cdot}400Y + 0{\cdot}375Z.$$

The unit equation therefore represents the "proportions" of a colour rather than its "amount "or relative luminosity.

The relation between unit equations of a colour with different primaries may also cause difficulty to a student, and a diagram such as figure 9 may help in the understanding of the transformation equations. This diagram shows the unit triangle of the *RGB* primaries of the author's colorimeter (Holmes, 1935) superimposed on the usual rectangular *XYZ* diagram in such a way that the unit *RGB* equation and the unit *XYZ* equation of any colour both plot at the same point in the two scales. For example, the unit equations for Illuminant B plot at:—

$S_B = 1/3 \,.\, R + 1/3 \,.\, G + 1/3 \,.\, B$ in the *RGB* triangle, and
$S_B = 0{\cdot}348X + 0{\cdot}353Y + 0{\cdot}300Z$ in the *XYZ* triangle.

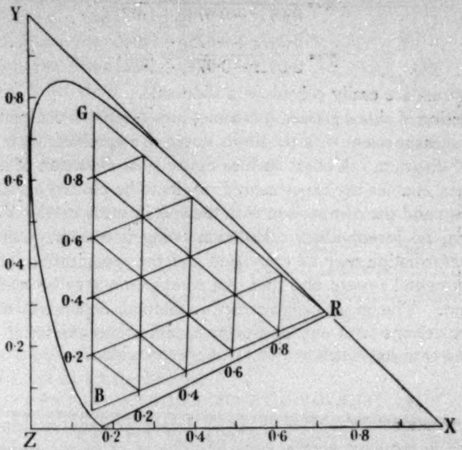


Figure 9. Relation between unit *RGB* triangle and unit *XYZ* triangle.

The *RGB* triangle has straight sides and it may be sub-divided by straight lines, although the scales are not quite evenly spaced. If monochromatic stimuli were employed (Smith and Guild, 1931, page 78) the *RGB* triangle would be nearly equilateral, but would have less evenly spaced scales for the coefficients of the three primaries.

The transformation equations to be employed in producing a chart such as figure 9 may be derived as follows:—

The calibration of the colorimeter gives three equations for the instrumental primaries:

$$R = x_1X + y_1Y + z_1Z,$$
$$G = x_2X + y_2Y + z_2Z,$$
$$B = x_3X + y_3Y + z_3Z.$$

In these equations, x_1 etc. are the coefficients determined by the calibration and usually the sum of all nine coefficients is 3·000.

A colour C can be represented by unit equations:

$C = rR + gG + bB$ on the instrumental (RGB) system.

$C = xX + yY + zZ$ on the C.I.E. (XYZ) system.

These equations may be reduced to a form which is easy to handle arithmetically:

$$x = \frac{r(x_1 - x_3) + g(x_2 - x_3) + x_3}{r[(x_1 + y_1 + z_1) - (x_3 + y_3 + z_3)] + g[(x_2 + y_2 + z_2) - (x_3 + y_3 + z_3)] + (x_3 + y_3 + z_3)},$$

$$y = \frac{r(y_1 - y_3) + g(y_2 - y_3) + y_3}{r[(x_1 + y_1 + z_1) - (x_3 + y_3 + z_3)] + g[(x_2 + y_2 + z_2) - (x_3 + y_3 + z_3)] + (x_3 + y_3 + z_3)}.$$

For the author's colorimeter, the transformation equations are:

$$x = \frac{0{\cdot}594r + 0{\cdot}001g + 0{\cdot}150}{0{\cdot}011r - 0{\cdot}074g + 1{\cdot}021},$$

$$y = \frac{0{\cdot}242r + 0{\cdot}675g + 0{\cdot}046}{0{\cdot}011r - 0{\cdot}074g + 1{\cdot}021}.$$

These equations are easily solved on a slide rule.

In the testing of signal glasses, it is sometimes desired to compare the result of a colorimeter measurement with the limits stated in a specification in terms of areas on the XYZ diagram. A chart such as figure 9, or a portion of it, drawn on an enlarged scale, enables the experimental results to be plotted on the instrumental (RGB) system and the comparison with the specification on the XYZ system can then be seen, no intermediate calculation being necessary. Alternatively, the reverse transformation may be calculated and the specification limits converted to the instrumental system and then the comparison may be made on a simple RGB diagram. The choice between the two methods depends on the frequency of making comparisons with any particular specification, the latter method being preferable for routine work.

§ 6. THE DESIGN OF COLOUR FILTERS

The requirements for a glass colour filter may be stated in terms either of the colour which is required with a given illuminant or of the spectrophotometric transmission curve. The former statement of the problem is usually easier to satisfy, because two glasses can often be found on opposite sides of the required colour and a match can be obtained by subtractive mixing. As an example of this, the Aviation Green glass in figure 7 was obtained by a mixture of copper oxide and chromium oxide, data for each of which are shown in figure 4. In obtaining the match, an hour or so of calculation by trichromatic methods may save days of experimental meltings. Allowance must be made for the base glass employed, that is to say, for the network and modifiers into which the colouring ions are to be admitted, for the conditions of oxidation or reduction in the melting process for traces of impurities or of oxides which react with the colouring oxides and then the degree of difficulty depends on how closely the required colour has to be matched.

If a transmission curve has to be matched, the problem is more complex. The transmission curves of most of the simple colouring oxides on the simple base glasses are known and some laborious arithmetic may enable a fairly close match to be obtained by subtractive mixing of two or three transmission curves. The arithmetic is greatly simplified if the curves are in terms of internal density, when the process of subtractive mixing is one of simple arithmetical addition at each wave-length, rather than multiplication. If the calculation is done graphically, the use of proportional dividers for scaling from a curve of internal density (or of transmission on a logarithmic scale) and wave-length enables any given thickness or concentration of each colouring oxide to be used in the calculation. Having obtained an approximate answer, you try it, and by successive approximations, gradually approach your target. There is scope for considerable personal skill in choosing the right starting point and estimating the corrections to be applied.

As an example of what can be done, figure 10 shows the transmission of some colour-temperature conversion filters; the curved dashed line represents the type of glass filter generally available in this country before the war, and the straighter line represents a filter which has been developed during the war. This diagram is plotted on an unusual scale suggested by Gage (1933) with the transmission on a logarithmic scale and the wave-length on a reciprocal scale, chosen because a perfect colour-temperature conversion filter, based on the Wien equation, would plot as a straight line.

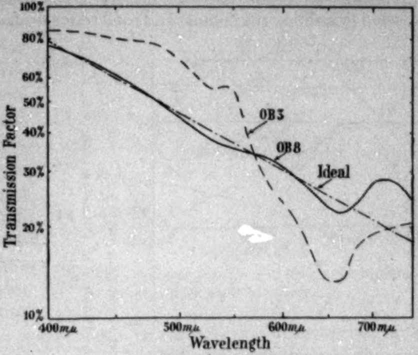


Figure 10. Transmission curves for colour-temperature conversion filters.

§ 7. THE COLOURS OF BLOOMED LENSES AND OF POLARIZED LIGHT

It is usual to estimate the reflexion factor of a bloomed surface, as used on a lens in an optical instrument, by looking at its colour; and the relation between these two provides a rather interesting use of colorimetry. If a film of low index is put on the surface of a glass, the combined reflexion factor of the two surfaces of the film is less than that of the single surface of the glass and, if the film is thin enough for destructive interference to take place, the reflexion factor can be reduced very considerably at some part of the spectrum. The reflected light is therefore coloured and the colour and the total reflexion factor are related to each other. Figure 11 shows the curves of reflexion factor calculated by classical theory for different values of the optical thickness (nt) of a film of refractive index (n) 1·30 on a base glass of refractive index 1·69. The reflexion factor of one surface of the base alone is 6·8% and the addition of the film would reduce this to 3·4% if no interference took place. If the interference is destructive, as for a wave-length of 0·54 μ and an optical thickness of 0·135 μ, the reflexion factor is zero, but if the interference is reinforcing, as for a wave-length of 0·50 μ and an optical thickness of 0·25 μ, the reflexion factor is 6·8%. The curves in figure 11 show that thin films ($nt = 0{\cdot}05\,\mu$ or $0{\cdot}10\,\mu$) are yellow-orange-red by reflected light, that a film of about $0{\cdot}135\,\mu$ optical thickness will have a low total reflexion factor and a purple colour and that thicker films ($nt = 0{\cdot}2\,\mu$ or $0{\cdot}25\mu$) are blue-green by reflected light.

If the trichromatic coefficients and total reflexion factors are calculated for a number of thicknesses and plotted as in figure 12 it will be seen that the minimum total reflexion factor occurs at about $0{\cdot}14\,\mu$ for the optical thickness (nt) of the film and that there is a very rapid change in the trichromatic coefficients near this thickness. It follows that there can be a wide variation in colour without any significant variation in reflexion factor. This is also shown in figure 13, where the solid line shows the colours and total reflexion factors calculated from curves as in figure 11 and the dashed line shows the colours of less saturation which are obtained if the interference is not completely destructive as, for example, when the refractive index of the film is not the square root of that of the base glass. Considering the solid line, a reflexion factor of 0·3% may be obtained with a red-coloured (thin) film or a blue-coloured (thick) film and it follows that films of intermediate colours will have lower reflexion factors. The same general conclusions may be applied to the dashed line, which shows the colours likely to be obtained from real films in which the coloration produced by destructive interference is diluted by white light.

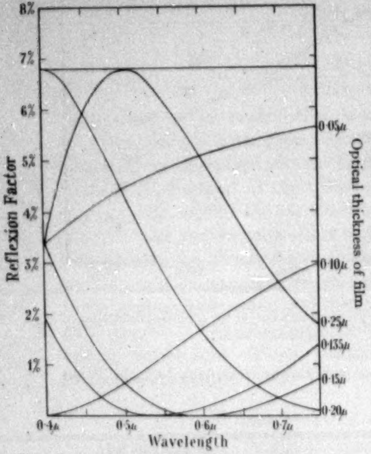


Figure 11. Calculated reflexion factor curves for bloomed glass. (Several film thicknesses.)

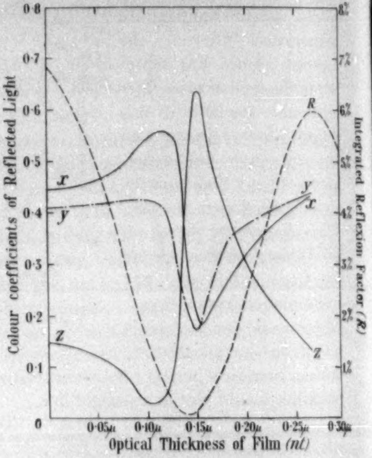


Figure 12. Calculated relation between trichromatic coefficients and film thickness.

Figure 13 is very similar to the diagram in Dr. Wright's *The Measurement of Colour* (1945), based on the work of Baud and Wright (1930) in which photoelastic colours are illustrated. These colours are obtained by an entirely different physical process but it so happens that both interference and polarization give a cosine law for the intensity distribution and the chromaticity diagrams are therefore similar.

This leads to another use of colour in the manufacture of glass, namely the inspection of glass for internal strain by polarized light.

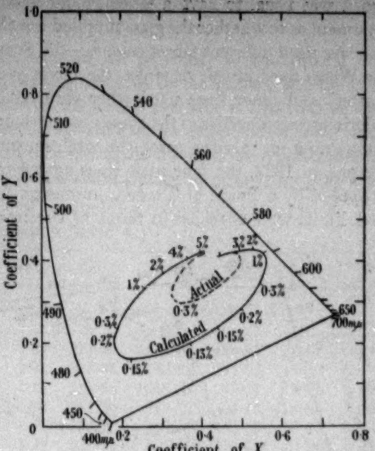


Figure 13. Colours and reflexion factors for bloomed glass with 2848° K. source.

If the retardation is of the order of a wave-length, there is a very rapid change in colour for a relatively small change in retardation or, because the retardation in a strained glass is proportional to the stress, for a relatively small change in stress. This is similar to the rapid change in colour for a small change in film thickness on bloomed lenses. In a strain-viewer or polariscope for examining the annealing of glass, a tint plate is used, which has one wave-length retardation for yellow-green light ($0{\cdot}55\,\mu$), and a small change in this retardation due to a small strain in the glass will change the colour of the transmitted light and, as it appears that the changed colour lies actually in the strained glass, the regions of maximum strain are readily found and the amount of this strain may be estimated.

Opinions differ as to the best colour to use for the tint plate—whether it should be $0{\cdot}54\,\mu$ or $0{\cdot}57\,\mu$ retardation—but actually this is a simple problem in colorimetry and capable of an exact solution. The diagram given by Dr. Wright shows the colours given by different retardations with a light source at about 3000° K. and the scale of retardation is most widely spaced at about $0{\cdot}56\,\mu$. If daylight were used, the effective wave-length of the white light would be less and $0{\cdot}55\,\mu$ or $0{\cdot}54\,\mu$ might be better. The exact optimum value can be obtained by calculating the colours with the particular light source involved and plotting them on a uniform chromaticity scale (Holmes, 1940) as in figure 14. The optimum retardation is at the point on the curve where a change in retardation causes the maximum linear displacement of the colour as plotted, and the answer from Baud and Wright's work is $0{\cdot}572\,\mu$. This is one of the uses to which a uniform chromaticity scale may be put with complete confidence because small departures from true uniformity will not affect the result by an appreciable amount.

§ 8. THE SPECIFICATION OF COLOURED GLASSES

The specification of coloured glasses has changed very considerably during the last twenty years, and the story of glasses for coloured light signals can be used as an example of the developments which have taken place. Over twenty-five years ago it was usual to have a target colour and then to have a certain amount of argument as to whether the glass supplied was the same as the target. The various railways used different target colours—the Scottish Railways used emerald green, the Welsh used purple or violet, the Great Western used blue-green (with limit glasses) and other lines used other varieties of green. Our knowledge of the colorimetric properties of the glasses was no more than the end-point of a spectrum photograph or, in a few cases, the total transmission factor. When the railways merged in 1922, the four main lines agreed to adopt the red and green glasses proposed by a Board of Trade Commission for ships' lights, and the light and dark limits were specified in terms of spectrophotometric transmisssion curves.

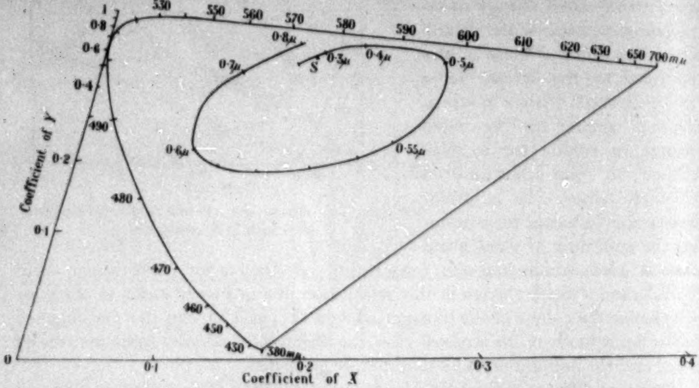


Figure. 14. Colours and retardation for polarization colours (after Baud and Wright).

The glasses shown were well suited to oil flame illuminants and limit glasses were specially made and precisely specified. In 1928 an orange range was introduced, the choice of the colour being based on laboratory experiments at the N.P.L. (Guild, 1928) on the risk of confusion with red or with white signals. Limit glasses were chosen and recorded in terms of their spectrophotometric transmission curves.

In 1933 the British Standards Institution set up a representative committee of Signal Engineers of Railways, representatives of the Ministry of Transport, Signal Makers, National Physical Laboratory, and others, to prepare a specification for coloured railway signal glasses and full use was made of the trichromatic methods of description of colour which had been agreed by the C.I.E. in 1931. A series of full-scale tests showed which glasses were too dark or too light for each colour and for each type of signal and the limiting glasses were measured with a standard light source (2360° K.) at the N.P.L. The colours were plotted on a chromaticity chart and it was possible to prescribe the areas on the chart within which a glass must lie if it was to be acceptable in service. The specification

BSS 623 was issued in 1935 and was revised in 1940, when only slight changes were necessary.

In 1937 laboratory-scale experiments on the reliability of recognizing any particular colour as red or green had been commenced by the author; and the 1940 revision of the specification was based on the results available at that time as well as on full-scale experience. These colour-recognition experiments have since been completed (Holmes, 1941) and a new specification for coloured glasses for all types of signal—railway, marine, avaition, street traffic—is now being prepared,

Table 1. Abstract from table of products for colour calculation with Illuminant A (2848° K.)

Wavelength (mμ)	400	450	500	550	600	650	700	750
Transmission factor	Products $T.E.\bar{x}$ to give coefficient of X							
100%	1·93	103·20	2·69	373·29	1271·03	434·47	20·67	0·63
90	1·74	92·88	2·42	335·96	1143·92	391·02	18·60	0·57
80	1·54	82·56	2·15	298·63	1016·82	347·58	16·54	0·50
70	1·35	72·24	1·88	261·30	889·72	304·13	14·47	0·44
60	1·16	61·92	1·61	223·97	762·62	260·68	12·40	0·38
50	0·97	51·60	1·35	186·65	635·52	217·24	10·34	0·32
40	0·77	41·28	1·08	149·32	508·41	173·79	8·27	0·25
30	0·58	30·96	0·81	111·99	381·31	130·34	6·20	0·19
20	0·39	20·64	0·54	74·66	254·21	86·89	4·13	0·13
10	0·19	10·32	0·27	37·33	127·10	43·45	2·07	0·06
Wavelength (mμ)	400	450	500	550	600	650	700	750
Transmission factor	Products $T.E.\bar{y}$ to give coefficient of Y (Divide sum by 20 to give total transmission for 2848° K.)							
100%	0·05	11·67	179·56	857·07	754·60	163·89	7·44	0·21
90	0·04	10·50	161·60	771·36	679·14	147·50	6·70	0·19
80	0·04	9·34	143·65	685·66	603·68	131·11	5·95	0·17
70	0·04	8·17	125·69	599·95	528·22	114·72	5·21	0·15
60	0·03	7·00	107·74	514·24	452·76	98·33	4·46	0·13
50	0·02	5·84	89·78	428·54	377·30	81·94	3·72	0·10
40	0·02	4·67	71·82	342·83	301·84	65·56	2·98	0·08
30	0·02	3·50	53·87	257·12	226·38	49·17	2·23	0·06
20	0·01	2·33	35·91	171·41	150·92	32·78	1·49	0·04
10	0·00	1·17	17·96	85·71	75·46	16·39	0·74	0·02
Wavelength (mμ)	400	450	500	550	600	650	700	750
Transmission factor	Products $T.E.\bar{z}$ to give coefficient of Z							
100%	9·16	543·91	151·32	7·49	0·96	0·00	0·00	0·00
90	8·24	489·52	136·19	6·74	0·86	0·00	0·00	0·00
80	7·33	435·13	121·06	5·99	0·77	0·00	0·00	0·00
70	6·42	380·74	105·92	5·24	0·67	0·00	0·00	0·00
60	5·50	326·35	90·79	4·49	0·58	0·00	0·00	0·00
50	4·58	271·95	75·66	3·75	0·48	0·00	0·00	0·00
40	3·66	217·56	60·53	3·00	0·38	0·00	0·00	0·00
30	2·75	163·17	45·40	2·25	0·29	0·00	0·00	0·00
20	1·82	108·78	30·26	1·50	0·19	0·00	0·00	0·00
10	0·92	54·39	15·13	0·75	0·10	0·00	0·00	0·00

using this work and some parallel experiments on aviation signals made in 1939 at the Royal Aircraft Establishment (Hill, 1939). The method of preparing the specification is to start with the data from practical experience, record it in trichromatic terms and to use the results of laboratory experiments to rule out the abnormal or irrelevant data, to interpolate or to extrapolate and to make deductions from the practical data. The conclusions can be expressed in terms of the properties of the satisfactory glasses when measured with a standard light source (Illuminant A—2848° K.) by a standard method and recorded on a chromaticity diagram. The correlation of the data presents a most intriguing problem in the interpretation of colour recognition as well as in the calculation of colour and of the change of colour with changed conditions; and the work before this B.S.I. committee is one of the most promising possibilities of advance of colorimetric technique in the glass industry.

REFERENCES

Baud, R. V. and Wright, W. D., 1930. *J. Opt. Soc. Amer.*, **20,** 381.

British Standard Specification, 623–1940. "Colours for Signal Glasses for Railway Purposes."

Gage, H. P., 1933. *J. Opt. Soc. Amer.*, **23,** 46; 1937. *Ibid.*, **27,** 160.

Guild, J., 1928. *Proc. Int. Conf. Illum.*, p. 862.

Hardy, A. C., 1936. *Handbook of Colorimetry* (Mass. Inst. Tech., Cambridge, Mass., U.S.A.).

Hill, N. E. G., 1938. *Report No. E & I* 1159.

Holmes, J. G., 1935. *Proc. Phys. Soc.*, **47,** 400; 1937. *Proc. Inst. Lighthouse Conf. Berlin*, pp. 99–101 (German text) and pp. 77–79 (English text); 1940. *Proc. Phys. Soc.*, **52,** 359; 1941. *Trans. Illum. Engng. Soc., London*, **6,** 71.

MacAdam, D. L., 1935. *J. Opt. Soc. Amer.*, **25,** 361.

McLeod, J. H., 1945. *J. Opt. Soc. Amer.*, **35,** 185.

Morey, G. W., 1938. *The Properties of Glass* (New York: Reinhold Publishing Corporation), p. 34.

Powell, H. E. 1945. *J. Opt. Soc. Amer.*, **35,** 428.

Scholes, S. R., 1945. *Glass Ind.*, **26,** 417.

Sharp, D., 1942. *Glass Ind.*, **23,** 331.

Smith, T., 1934. *Proc. Phys. Soc.*, **46,** 372.

Smith, T. and Guild, J., 1931. *Trans. Opt. Soc.*, **33,** 73.

Vaughan, T. C., 1944. *Glass Ind.*, **25,** 259.

Weyl, W. A., 1944. *J. Soc. Glass Tech.*, **28,** 158.

Wright, W. D., 1945. *The Measurement of Colour* (London: Adam Hilger Ltd.), p. 196.

ULTRA-VIOLET BANDS OF Na$_2$

By S. P. SINHA,

Imperial College of Science and Technology

MS. received 4 *December* 1946

ABSTRACT. The ultra-violet bands of Na$_2$ have been studied in absorption and emission. In absorption, by varying the conditions of temperature and pressure the bands have been photographed from λ3640 A. to λ2500 A. They are considered to belong to seven different systems, of which, however, only three are well developed. The classification of the other four systems is tentative. It has been shown that the Na$_2$ ultra-violet bands measured by Walter and Barratt (1928), which were analysed into five different systems by Weizel and Kulp (1930), may really be considered to belong to two systems only, corresponding to systems 1 and 3 of the present investigation, which are the most intensely developed ones.

The emission bands are weakly developed and have been observed only in the regions $\lambda\lambda$3370–3180 A. and $\lambda\lambda$3070–2960 A. These correspond to the strong bands of systems 1 and 3 in absorption.

§1. INTRODUCTION

THE diatomic molecule of sodium is known to possess two systems of bands in the visible region, one of which lies in the yellow-red and the other in the green. Extensive studies of these two systems have been made by several authors: the vibrational structure of the yellow-red bands by Fredrickson and Watson (1927) and Fredrickson and Stannard (1933), their rotational structure by Fredickson (1929); the vibrational structure of the green bands by Loomis and Nusbaum (1932), and their rotational structure by Loomis and Wood (1928). The vibrational and rotational constants of the molecule for the states involved in these two systems of bands are thus known to a high degree of accuracy.

The sodium molecule is known also to possess some bands in the ultra-violet region in absorption. These were first observed by Wood (1909). Walter and Barratt (1928) also obtained these bands, and their measurements were arranged into five different systems by Weizel and Kulp (1930). Further investigations in this region were carried out by Kimura and Uchida (1932) who, using the light from the crater of a carbon arc, photographed the absorption spectrum of sodium on a Hilger E quartz spectrograph and arranged the bands thus obtained into as many as six systems. Although Kimura and Uchida attributed a larger number of bands to each system than did Weizel and Kulp, and since they also obtained a large number of bands at shorter wave-lengths not observed by Walter and Barratt, some of their systems appeared to be rather incomplete, and it seemed worth while to photograph the bands and attempt a new analysis. Accordingly further observations have been made both on the absorption and emission spectra. The results of measurements and conclusions are described in the following sections.

§2. SOURCE, APPEARANCE OF THE SPECTRUM AND MEASUREMENTS

(a) *Bands in absorption*

The details of the apparatus used in the absorption experiment during the present investigation have been described elsewhere (Bhattacharya and Sinha, 1943). Sodium, freed from the oil in which it was stored, was put in a steel cell kept inside an electrically heated steel tube provided with water-cooled quartz windows, and light from a hydrogen discharge tube was used as source for the continuous radiation. The hydrogen gave a perfect continuum except for some OH bands near λ3100 A., which were eliminated during measurements. The pressure inside the absorption tube could be varied by introducing nitrogen gas from a cylinder. The presence of nitrogen inside the absorption chamber has been found to facilitate the appearance of Na_2 bands in the ultra-violet region.

Preliminary investigations were made with a Hilger Intermediate quartz spectrograph, and the optimum conditions of temperature and pressure for the bands to develop satisfactorily in different regions were noted. Using these values of temperature and pressure, the spectrum was next photographed in an E_1 quartz instrument.

At about 705° c., when the total pressure in the chamber is about 5 cm. of mercury, the bands are strongest in the region from λ 3280 A. to λ 3450 A., and the λ 3303 A. line of sodium is not very broad. On increasing the temperature and pressure, nearly the whole of this region is continuously absorbed and bands appear at wave-lengths greater than λ 3450 A. Bands also appear quite strongly between λ 2880 A. and λ 3100 A., and fainter bands extend up to about λ 2500 A. The conditions that favour the satisfactory development of bands in different regions are:

Temperature (°c.)	750	850	900
Total pressure in the tube (cm. Hg)	5	15	25
Region	λλ 3280–3450 A.	λ 3400–3640 A. λ 2880–3280 A.	λ 2500–2880 A.

and measurements included in table 1 in different regions correspond to these values of temperatures and pressures.

Measurements were made of all the band-heads obtained on both the spectrographs. Those measured on the smaller dispersion instruments are, however, not presented here except in the regions λ 3639 A. to λ 3542 A. and λ 2837 A. to λ 2496 A. in which regions bands on the spectrograms obtained on E_1 were too faint to be measured.

Bands were nearly all degraded to the red, although in some cases the heads were not very prominent under the microscope. In some cases it was difficult to decide which way the band was degraded and in such cases (marked with asterisks) the centres of absorption were measured. Comparison of readings made with different spectrograms on the E_1 instrument showed that the measurements could be relied upon up to ± ·3 A. in most cases, although in some they were uncertain up to as high as ± ·5 A. Most of the bands measured could be classified and are given in table 1. The few remaining ones were mostly faint. Intensities given in table 1 are visual estimates on a scale of 10, and have been taken from different spectrograms for the different regions separated by horizontal lines in table 1.

(b) *Bands in emission*

The main purpose of studying the ultra-violet bands of sodium in emission was to settle the problem of arranging them into different systems.

The source was a discharge tube of simple design and essentially not different from that used by Wood and Galt (1911) and Kimura and Uchida (1932) for studying the visible and the ultra-violet bands of Na_2 respectively. The tube was of Pyrex glass 50 cm. long and 3·5 cm. in diameter and having three side tubes, two to carry the electrodes and the third to be connected to the vacuum pump. One end of the tube was closed and, after introducing some sodium freed from oil in the centre of the tube, a quartz window was sealed to the other end. The central portion of the tube over a length of about 12 cm. could be heated electrically from outside up to about 500° c. No cooling near the window was necessary.

The tube was first excited by an induction coil and the spectrum was photographed for different potentials over a range corresponding to about 4 to 10 inches of spark gap in air. The temperature was between 300° and 500° c. Although the visible bands appeared, there was no trace of any band in the ultra-violet. The tube was then excited by a transformer capable of giving a large current density though not a large potential. Under this excitation, at about 500° c., when the pressure in the tube was about 1 mm. of mercury, a brilliant yellow light appeared. The visible bands could be photographed quite readily, and in less than an hour's exposure some ultra-violet bands also appeared on the plate with the medium quartz instrument.

The measurements are given in table 2. The bands appear in three distinct regions: (1) λ 3370 A. to λ 3180 A., (2) λ 3070 A. to λ 3000 A. and (3) λ 2960 A. to λ 2900 A. In appearance they resemble the absorption bands on the Intermediate quartz spectrograph, except that the latter are much more intense.

Table 1. Ultra-violet absorption bands of Na_2

λ_{air} (A.)	Intensity	v', v''	System
3639	1	2,16	1
3619·5	1	2,15	1
3600	1	1.13	1
3580	2	1,12	1
3561·5	2	0,10	1
3542	2	0,9	1
3510·2	2	1,8	1
3505·4	2	0,7	1
3479·6	2	2,7	1
3474·6	3	1,6	1
3452·9	4	4,7	1
		0,4	1
3448·7	2	3,6	1
3443·5	2	2,5	1
3439·0	5	1,4	1
3434·8	6	4,6	1
3424·8	2	2,4	1
3416·4	4	0,2	1
3412·5	2	3,4	1
3409·7	1	6,6	1
3408·7	2	2,3	1
3402·7	3	1,2	1
3399·7	1	4,4	1
3397·7	6	0,1	1
3393·8	2	3,3	1
3389·7	3	2,2	1
3384·7	8	1,1	1
3381·1	1	4,3	1
3374·3	1	6,4	1
3368·5	1	5,3	1
3366·5	6	1,0	1
3358·8	4	3,1	1
3355·8	1	6,3	1
3352·9	8	2,0	1
3342·6	1	7,3	1
3340·6	10	3,0	1
3337·4	2	6,2	1
3333·9	4	5,1	1
3327·9	10	4,0	1
3325·2	1	3,15	2
3316·0	8	5,0	1
3312·7	2	4,15	2
3309·1	4	7,1	1
3304·1	3	6,0	1
3298·9	4	8,1	1
		4,14	2
3295·8	2	3,13	2
3291·8	5	7,0	1
3287·3	1	5,14	2
3283·0	1	4,13	2
3280·4	4	8,0	1
		3,12	2
3278·5	1	7,15	2
3274·6	4	10,1	1
		6,14	2
3271·6	2	5,13	2
3269·3	4	9,0	1
3268·2	1	4,12	2
3266·9	1	8,15	2
3265·8	2	3,11	2
3264·2	3	11,1	1
3261·7	1	2,10	2
3259·9	1	6,13	2
3258·6	1	13,2	1
3257·9	2	10,0	1
		1,9	2
3257·0	2	5,12	2
3253·3	3	12,1	1
		4,11	2
3248·9	1	7.13	2

Table 1. Ultra-violet absorption bands of Na_2 (*cont.*)

λ_{air} (A.)	Intensity	v',v''	System
3247·9	1	14,2	1
3246·2	2	2,9	2
3242·4	3	13,1	1
		5,11	2
		1,8	2
		4,14	2
3238·3	2	4,10	2
3236·5	2	12,0	1
3235·2	2	3,9	2
3231·8	2	14,1	1
3227·4*	2	16,2	1
		5,10	2
		1,7	2
3225·6	1	13,0	1
3223·6	2	0,6	2
		4,9	2
3221·7	1	15,1	1
3219·5	2	3,8	2
3217·7	1	17,2	1
3215·5	2	2,7	2
3211·9	1	16,1	1
3207·0	1	18,2	1
3204·8	2	3,7	2
3202·1	1	17,1	1
3200·0	2	2,6	2
3196·7	3	19,2	1
		1,5	2
3191·5	1	21,3	1
3187·8	1	20,2	1
3181·4	2	1,4	2
3178·2	2	0,3	2
3176·4	1	21,2	1
3170,8	2	2,4	2
3160·0	2	3,4	2
3156·2	2	2,3	2
3151·6	2	1,2	2
3145·2	2	3,3	2
3140·0	2	2,2	2
3135·7	3	1,1	2
3131·2	2	0,0	2
3125·1	2	2,1	2
3120·5	2	1,0	2
3118·0	2	1,8	3
3110·6	1	3,9	3
3109·5	2	2,0	2
3107·4	1	2,8	3
3103·4	1	4,1	2
3098·6	1	3,0	2
3097·0	3	6,2	2
		3,8	3
3092·9	3	5,1	2
		2,7	3
3089·8	3	5,9	3
3088·2	3	4,0	2
		1,6	3
3086·6	2	6,2	2
		13,6	2
		4,8	3
3082·2	3	6,1	2
		3,7	3
3078·0	3	5,0	2
		3,6	3
3076·3	4	8,2	2
		11,4	2
		5,8	3
3074·1	3	1,5	3
3072·6	1	13,5	2
3071·6*	2	7,1	2
		10,3	2
		4,7	3
3069·6	1	7,9	3
3067·6	3	3,6	3
		6,0	2
		12,4	2
3060·1	4	1,4	3
3057·7*	2	13,4	2
		7,0	2
3056·6	2	10,2	2
3055·6	4	0,3	3
3054·5	1	12,5	2
3052·8	1	12,3	2
3052·2	2	9,1	2
3049·0	1	14,4	2
		2,4	3
3047·6	4	8,0	2
		11,2	2
3046·0	1	16,5	2
3045·6	4	1,3	3
3043·6	1	13,3	2
3041·7	1	0,7	4
3041·0	4	0,2	3
3039·0	1	3,9	4
3034·8	3	14,3	2
		2,3	3
3031·7	3	1,7	4
3031·0	5	1,2	3
3027·9	3	0,6	4
3027·2	5	0,1	3
3026·0	2	3,3	3
3021·0	2	2,2	3
3018·5	3	1,6	4
3016·9	5	1,1	3
3014·7	2	0,5	4
3012·1	5	0,0	3
3009·0	4	2,6	4
3007·0	4	2,1	3
3004·7	2	1,5	4

Table 1. Ultra-violet absorption bands of Na_2 (*cont.*)

λ_{air} (A.)	Intensity	v',v''	System	λ_{air} (A.)	Intensity	v',v''	System
3002·3	5	1,0	3	2886·7	2	17,2	3
3000·3	2	4,0	4	2884·5	1	7,0	4
2995·7	4	2,5	4	2884·0	1	19,3	3
2992·6	5	2,0	3	2882·4	1	12,3	4
2990·9	3	1,4	4	2880·1	2	9,1	4
2987·9	4	4,1	3	2877·4	2	11,2	4
2986·4	6	0,3	4	2876·3	1	8,0	4
2983·1	6	3,0	3	2874·0	1	13,3	4
2978·2	4	5,1	3	2872·8	1	10,1	4
2977·0	5	1,3	4	2869·8	1	12,2	4
2973·8	4	4,0	3	2866·5	1	14,3	4
2972·0	4	0,2	4				
2968·6	5	6,1	3	2837	1	0,2	5
2964·6	5	5,0	3	2829	1	1,2	5
2963·8	4	8,2	3	2824·5	3	0,1	5
2959·6	6	7,1	3	2816	1	1,1	5
2958·6	5	0,1	4	2808	2	2,1	5
2955·2	6	6,0	3	2800	1	3,1	5
2953·7	4	9,2	3	2792	2	4,1	5
2949·7	6	8,1	3	2780·5	2	4,0	5
		1,1	4	2773	2	5,0	5
2948·2	6	2,1	4	2765	3	6,0	5
2945·5	10	7,0	3	2758	3	7,0	5
2944·0	5	0,0	4	2750	5	8,0	5
2941·6	3	12,3	3	2746	2	0,7	6
2936·3	8	8,0	3	2742	1	9,0	5
		11,2	3	2738	2	1,7	6
2935·6	8	1,0	4	2735	5	0,6	6
2932·5	6	10,1	3	2730	4	2,7	6
2931·7	3	3,1	4	2727	3	1,6	6
2928·6	6	12,2	3	2719	2	2,6	6
2927·6	8	9,0	3	2701	3	3,5	6
2920·6	5	2,0	4	2690	2	3,4	6
2924·4	3	11,1	3	2672	1	4,3	6
2922·9	3	4,1	4	2665	2	5,3	6
2920·3	5	13,2	3	2658	1	6,3	6
2919·1	3	10,0	3	2654	1	5,2	6
2917·2	3	3,0	4	2651	1	7,3	6
2915·8	5	12,1	3	2647	1	6,2	6
2914·7	4	5,1	4	2643·5	1	5,1	6
2912·9	3	17,4	3	2640·5	1	7,2	6
2912·0	3	14,2	3	2637	1	6,1	6
2908·8	3	4,0	4	2630	1	7,1	6
2907·4	2	16,3	3	2623	1	8 1	6
2906·9	3	6,1	4	2617	1	9,1	6
2904·2	3	18,4	3	2612	1	8,0	6
2903·5	3	15,2	3	2605	1	9,0	6
2901·3	1	5,0	4	2598·5	1	10,0	6
2899·9	2	17,3	3	2577	1	0,4	7
2898·8	2	10,3	4	2570	1	1,4	7
2897·7	2	7,1	4	2563*	1	2,4	7
2896·0*	1	19,4	3	2552*	1	2,3	7
2892·9	1	6,0	4	2535*	1	3,2	7
2891·9*	1	18,3	3	2528*	1	4,2	7
2891·0*	1	15,1	3	2519*	1	4,1	7
2889·7	2	8,1	4	2503*	1	5,0	7
2888·4	1	20,4	3	2496*	1	6,0	7

Table 2. Na_2 ultra-violet bands in emission

λ_{air} (A.)	Intensity	v',v''	System	λ_{air} (A.)	Intensity	v',v''	System
3366	2	1,0	1	3190	1		
3358	4	3,1	1	3187	1	20,2	1
3353	4	2,0	1	3179	1		
3346	4	4,1	1				
3341	5	3,0	1	3073	2	1,5	3
3334	5	5,1	1	3068	1	0,4	3
3328	4	4,0	1	3059	2	1,4	3
3316	4	5,0	1	3045	4	1,3	3
3312	2			3040	4	0,2	3
3309	2	7,1	1	3031	5	1,2	3
3304	2	6,0	1	3026	5	0,1	3
3298	2	8,1	1	3021	5	2,2	3
3292	4	7,0	1	3017	5	1,1	3
3286	4	9,1	1	3012	5	0,0	3
3280	4	8,0	1	3008	5	2,1	3
3275	4	10,1	1	3002	4	1,0	3
3269	3	9,0	1	2997	2	3,1	3
3264	4	11,1	1	2959	2	7,1	3
3253	4	12,1	1	2950	2	8,1	3
3247	2	11,0	1	2941	4	9,1	3
3236	2	12,0	1	2939	4		
3229	2	16,2	1	2936	4	8,0	3
3226	2	13,0	1	2932	2	10,1	3
3221	4	15,1	1	2929	2		
3218	1	17,2	1	2927	4	9,0	3
3215	2	14,0	1	2923	2		
3207	1	18,2	1	2921	2	13,2	3
3202	1	17,1	1	2918	4	10,0	3
3197	1	19,2	1	2908	2		

§3. VIBRATIONAL ANALYSIS

The absorption bands measured in the present investigation have been found to belong to seven systems. Of these, only three systems, viz. 1, 2 and 3, appear at all extensive. The remaining four look like fragments, and until they are photographed under more favourable conditions, the genuiness of their separate existence will remain questionable.

The classification and assignment of quantum numbers have already been given in table 1. In addition, tables 3, 4 and 5 give the Deslandres schemes for the important portions of the three well-developed systems. The mean differences for the upper state given in these tables do not appear quite regular. Since the probable errors in measurements are mostly as high as ±0·3 A. and sometimes even higher than this, the irregularities shown in these differences may be due to experimental errors and do not seem worthy of any special investigation with regard to perturbation or the like phenomena unless the measurements are improved. A somewhat similar irregularity is also shown by the differences for the ground state, which is known to be free from any perturbation. A few bands appear in brackets in these tables; they have not been observed but have been obtained by interpolation, and have been utilized for calculating the differences

Table 3. Vibrational scheme for the first ultra-violet system of Na_2 bands

v'' / v'	0		1		2		3		4		5		6		7		8		9	Mean diff.
0					29262	155	(29107)	154	28953						28518	147	(28371)	146	28225	
					118		117		117						107		109			114
1	29696	160	29536	156	29380	156	(29224)	154	29070	149	(28921)	149	28772	147	(28625)	145	28480			
	120				113		114		120		111		110		106					113·5
2	29816				29493	155	29338	148	29190	158	29032	150	29882	151	28731					
	110				117		119		106		108		106							111
3	29926	162	29764	154	(29610)	153	29457	161	29296	156	(29140)	152	28988							
	114						110		111		116		117							113·5
4	30040						29567	160	29407	151	(29256)	151	29105	152	28953					
	108						111													109·5
5	30148	162	29986				29678													
	109																			109
6	30257																			
	113																			113
7	30370	158	30212																	
	105		102																	103·5
8	30475	161	30313																	
	104		108																	106
9	30579	157	(30422)																	
	107		107																	107
10	30686	157	30529																	
	102		98																	100
11	(30788)	161	30627																	
	101		102																	101·5
12	30889	160	30729																	
	104		103																	103·5
13	30993	161	30832																	
			102																	102
14			30934																	
			97																	97
15			31031																	
			94																	94
16			31125																	
Mean diff.		160		155		155		155·5		153·5		150·5		150		146		146		
Loomis & Nusbaum		(158)		(156)		(155)		(153·5)		(151)		(150·5)		(149)		(146·5)		(146)		

Table 4. Vibrational scheme for the second ultra-violet system fo Na_2 bands

v' \ v''	0		1		2		3		4		5		6		7		8		9		10		11	Mean diff.
0	31927						31453						31012											
	110												113											111·5
1	32037	155	31882	161	31721				31424	151	31273	148	31125	149	30976	144	30832	146	30686					
	113		108		115				105				116		114		110		110					111
2	32150	160	31990	154	31836	161	31675	146	31529				31241	151	31090	148	(30942)	146	30796	146	30650			
	112						110		98 (?)						104		110		105					108
3	32262						31785	158	31627	154	31473				31194	142	31052	151	30901					
	110																		111					110·5
4	32372	159	32213																31012	140	30872	143	30729	
	107		110																		104		103	106
5	32479	156	32323																		30976	144	30832	
	110		112																					111
6	32589	154	32435	155	32280																			
	106				108																			107
7	32695				32388																			
	108				109																			108·5
8	32803	153	(32650)	153	32497																			
			104		105																			104·5
9			32754	152	(32602)																			
					105																			105
10					32707	160	32547																	
					96		100																	98
11					32803	156	(32647)	150	32497	159	32536	148	32388											
							100		92		97													96
12							32747	158	32589	155	(32633)													
							99		106		96													102·5
13							32846	151	32695		32729													
							96		93		91													95
14							32942	154	32788		32820													
																								96
15																								
																								91
16																								
Mean diff.		156		155		159		153		155		148		150		145		148		143		143·5		
Loomis & Nusbaum		(158)		(156)		(155)		(153·5)		(151)		(150·5)		(149)		(146·5)		(146)		(144)		(143)		

Table 5. Vibrational scheme for the third ultra-violet system of Na_2 bands

v' \ v''	0		1		2		3		4		5		6		7		8		9	Mean diff.
0	33190	166	33024	150	32874	157	32717													
	108		113		109		107													109
1	33298	161	33137	154	32983	159	32824	155	32669	149	32520	148	32372				32062			
	108		110		109								107				109			109
2	33406	159	33247	155	33092								32479	156	32323	151	32172			
	106												110		112		110			109·5
3	33512												32589	154	32435	153	32282	149	32133	
	105												106		112					108
4	33617	158	33459										32695	148	32547					
	105		109																	107
5	33722	154	33568														32497	142	32355	
	107		108																	107·5
6	33829	153	33676																	
	111		103																	107
7	33940	161	33779																	
	106		112																	109
8	34046	155	33891																	
	102																			102
9	34148																			
	99																			99
10	34247	156	34091																	
			98																	98
11			(34189)																	
			97																	97
12			34286	150	34136															
			99		97															98
13			34385	152	34233															
			99		98															98·5
14			34486	155	34331															
			100		100															100
15			34586	155	34431															
			97		101															99
16			34683	151	(34532)	147	34385													
					100		91													95·5
17					34632	156	34476													
							93													93
18							34569													
Mean diff.		158		153		155		155		149		148		153		152		145·5		
Loomis & Nusbaum		(158)		(156)		(155)		(153·5)		(151)		(150·5)		(149)		(146·5)		(146)		

wherever only few bands for this purpose are available. The numbers given in brackets below the mean differences for the ground state are the values of the latter obtained from the work of Loomis and Nusbaum (1932). Further, the bands lie along a parabola whose shape is similar to what we can expect for the relative values of w_v'' and w_v' of the states involved in these systems.

Weizel and Kulp (1930), who have analysed the bands measured by Walter and Barratt (1928), consider the bands between λ 3370 A. and λ 3180 A. to belong to as many as three systems, while Kimura and Uchida (1932) consider the bands in the same region to belong to two systems. In the present scheme, however, all the intense bands in this region have been placed under system 1 only. It is also possible to accommodate all the bands due to Walter and Barratt in this region into a single system which corresponds to system 1 of the present arrangement. The reason why Weizel and Kulp considered them to belong to more than one system is that a few bands near λ 3303 A. are absent from the spectrogram, having been masked due to continuous absorption resulting from the broadening of the principal-series line. In the same way, systems 4 and 5 of Weizel and Kulp seem to be really another single system corresponding to system 3 of the present classification. System 4 of theirs corresponds to the right limb and system 5 to the left limb of the Condon parabola of the present system 3.

The classification of the emission bands is given in table 2. Bands between λ 3370 A. and λ 3180 A. correspond to the first u.v. system observed in absorption. The emission bands correspond to strong ones on the left limb of the Condon parabola. Bands between λ 3070 A. and λ 2900 A. correspond to the third system in absorption. Only the intense bands have appeared in emission. The second system observed in absorption, which is weaker than the first or the third system, has not appeared in emission. Bands between λ 3370 A. and λ 3600 A. (reported by Kimura and Uchida) are also absent.

The vibrational constants and heats of dissociation for the upper states of the three well developed systems are given in table 6. The heat of dissociation has been calculated by Birge and Sponer's extrapolation method, and its value thus suffers from the defects accompanying this method, especially when the range of extrapolation is large. The calculations for the dissociation products of these states shown in the last column of table 6 are given in the next section.

Table 6. Molecular constants for the upper states of the first three ultra-violet systems of Na_2 bands

Upper-state constants	$\nu_{0,0}$ (cm^{-1})	w_0' (cm^{-1})	$x_0'w_0'$ (cm^{-1})	D' (cm^{-1})	ν_{atom} (cm^{-1})	Dissociation products
System 1	29585	115	0·6	5500	28930	$3\,^2S+4\,^2P$ or $3\,^2S+3\,^2D$
System 2	31930	111	0·7	4400	30170	$3\,^2S+4\,^2P$
System 3	33190	109	0·5	5900	32530	$3\,^2S+5\,^2S$

§ 4. DISSOCIATION PRODUCTS

$\nu_{\text{atom}} = \nu_{0,0} + D' - D''$, where the terms used have their usual significance. Using values of $\nu_{0,0}$ and D' given in table 6, and assuming $D'' = 6160$ cm.$^{-1}$ Loomis and Nusbaum, 1932), we get the following values of ν_{atom}:

$$\nu_{\text{atom}}\,(\text{System 1}) = 28930 \text{ cm.}^{-1},$$
$$\nu_{\text{atom}}\,(\text{System 2}) = 30170 \text{ cm.}^{-1},$$
$$\nu_{\text{atom}}\,(\text{System 3}) = 32930 \text{ cm.}^{-1}.$$

Further, from the line-spectra data (Fowler's *Report*, 1922) it is known that for sodium

$$3\,{}^2S - 3\,{}^2D = 29160 \text{ cm.}^{-1},$$
$$3\,{}^2S - 4\,{}^2P = 30270 \text{ cm.}^{-1},$$
$$3\,{}^2S - 5\,{}^2S = 33200 \text{ cm.}^{-1}.$$

A comparison would thus indicate that the upper states of the first three systems of ultra-violet bands dissociate into a $3\,{}^2S$ and an excited $3\,{}^2D$, $4\,{}^2P$ and $5\,{}^2S$ atoms. We should then, however, expect two more systems of bands approximately in the same region as these systems, because each of the combination $3\,{}^2S$ and $3\,{}^2D$ or $3\,{}^2S$ and $4\,{}^2P$ is theoretically capable of giving rise to as many as four stable states, transitions to two of which from the ${}^1\Sigma_g{}^+$ ground state are permissible. There is no evidence for these systems, nor has any system been observed whose upper-state dissociation products could be $3\,{}^2S + 4\,{}^2S$ atoms of sodium. An equally likely dissociation product for the first system could be $3\,{}^2S + 4\,{}^2P$ atoms, for the discrepancy of about 1000 cm.$^{-1}$ would not be beyond the range of probable errors involved in calculating D'. Both the assignments for this state are therefore given in table 6. The assignments should, however, be considered merely tentative. These can be settled only if bands can be photographed at still higher dispersion, which might reveal members of higher v' values and also give the structure. Further work on this is in progress.

§ 5. ACKNOWLEDGMENTS

In conclusion, I wish to express my gratefulness to Assistant Professor R. W. B. Pearse, D.Sc., for his kind help and suggestions and to Professor D. K. Bhattacharya for his ungrudging help in the experimental work in absorption. My thanks are also due to the University of Patna, India, for the award of a scholarship which has enabled me to continue the work here.

REFERENCES

Bhattacharya, D. K. and Sinha, S. P., 1943. *Ind. J. Phys.*, **17**, 131.
Fowler, A., 1922. *Report on Series in Line Spectra* (London : Physical Society), p. 99.
Fredrickson, W. R., 1929. *Phys. Rev.*, **34**, 207.
Fredrickson, W. R. and Stannard, C. R., 1933. *Phys. Rev.*, **44**, 633.
Fredrickson, W. R. and Watson, W. W., 1927. *Phys. Rev.*, **30**, 429.
Kimura, M. and Uchida, Y., 1932. *Sci. Pap. Inst. Phys. Chem. Res.* (*Tokio*), **18**, 109.
Loomis, F. W. and Nusbaum, R. E., 1932. *Phys. Rev.*, **40**, 380.
Loomis, F. W. and Wood, R. W., 1928. *Phys. Rev.*, **32**, 223.
Walter, J. M. and Barratt, S., 1928. *Proc. Roy. Soc.*, A, **119**, 265.
Weizel, W. and Kulp, M., 1930. *Ann. Phys., Lpz.*, **4**, 971.
Wood, R. W., 1909. *Phil. Mag.*, **18**, 530.
Wood, R. W. and Galt, R. H., 1911. *Astrophys. J.*, **33**, 72.

EXPERIMENTS IN MULTIPLE-GAP LINEAR ACCELERATION OF ELECTRONS

By W. D. ALLEN and J. L. SYMONDS,
Radiophysics Laboratory, Sydney, N.S.W.

MS. received 6 *December* 1946

ABSTRACT. Using a CV76 magnetron at a wave-length of 10·0 cm. (maximum power 500 kw.; power available for acceleration, 300 kw.), and a 3-stage single cavity, electrons were accelerated to a voltage of 0·85 Mev. The expected figure was approximately 1·1 Mev. The paper describes the R.F. work involved in obtaining the result.

---

§ 1. INTRODUCTION

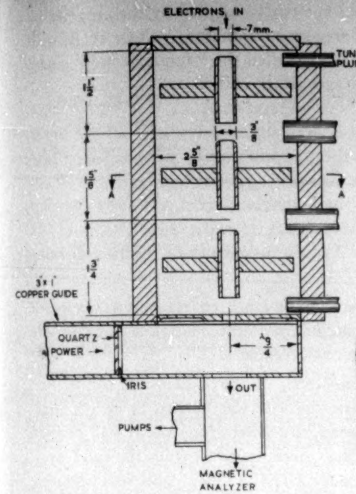
THE pioneer work on the linear accelerator of Lawrence and Sloan (1931) and Sloan (1935), indicated the possibility of obtaining fast particles with relatively low H.F. voltages, at frequencies of the order of 30 Mc./s. The linear accelerator has since been eclipsed by the development of the cyclotron: but its possibilities have again become prominent by the development of valves which furnish high-pulsed power at centimetre wave-lengths.

The use of 1200 Mc./s. power in a single cavity has been developed in this Laboratory and described (Bowen, Pulley and Gooden, 1946). To make optimum use of the power available, however, a series of, say, N gaps is necessary: since the power is then distributed between the gaps, the voltages per gap drops by N and the overall voltage increases by $\sqrt{N}$. There are various ways of doing this. A TM_{01} mode can be loaded with irises so that the phase velocity down the guide is equal to the particle (electron) velocity; so that the electron is carried forward continuously on the crest of the wave. The power can be fed into a series of re-entrant cavities, each fed separately, with suitable phasing from an arterial guide. Or the power can be fed into a single cavity, designed so that the voltages at consecutive gaps are suitably reversed. The last method was the one adopted in this work. It had the advantages that good bunching was achievable between first and second gaps: that by virtue of the tight coupling between separate sections, no drift of R.F. phase between the gaps was possible: and that the cavity acted as its own frequency stabilizer.

In what follows, a description is given of the cavity design, of phasing considerations, of high power work and of the acceleration experiments.

§ 2. CAVITY DESIGN

The essential features of the cavity are exhibited in figure 1 A. It consists in effect of a series of re-entrant cavities (figure 1 B) placed back to back, with a cut along the periphery of the dividing walls to provide the electrical coupling between the sections. At each end there is a half-section, in the one case for convenience of feeding and in the other of electron injection. The cut in the dividing wall necessitates the support of the centre " baffles "; this is accomplished by quarter-wave stub supports, as shown in the section view of figure 1 A.


ELECTRONS IN
7mm.
TUNING PLUNGERS
1½"
3/8"
2 5/8"
1 5/8"
A
A
1 3/4
3"×1" COPPER GUIDE
QUARTZ
POWER
λg/4
IRIS
OUT
PUMPS
MAGNETIC ANALYZER

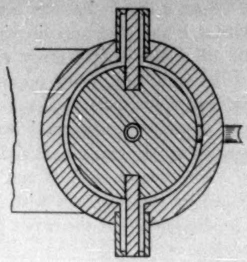


FIG. 1A. SECTION AA

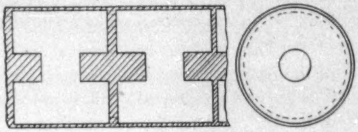


FIG. 1B.

Figures 1 A and 1 B.

The experiments on individual sections of the cavity were carried out on a cavity of the type shown in figure 2.1. This cavity may be regarded as two conventional re-entrant cavities, with gaps at B and C, coupled together by the annulus between disc and chamber wall at A. If we make the customary approximate representation of the re-entrant cavities as a resonant circuit

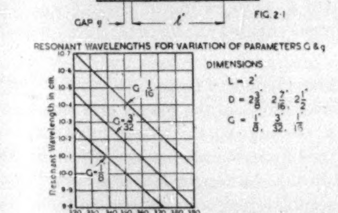

L'
A
GAP G'
C/L
D
B
C
2·601"
GAP g
l'

FIG. 2·1

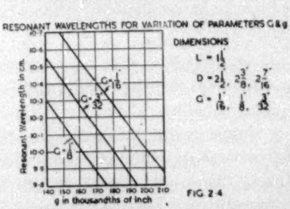


FIG. 2·2

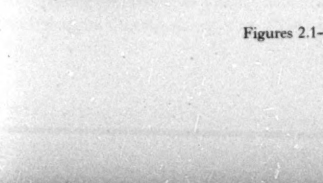


FIG. 2·3

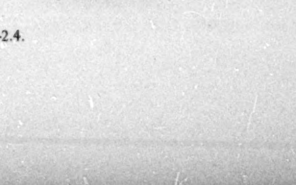


FIG. 2·4

Figures 2.1–2.4.

L and C_1 (figure 3), then we have two such circuits coupled by the capacity C_2 at A. Such a circuit has one resonant frequency with zero voltage at A, with voltages at B and C of opposite sign: this was called the "fundamental" mode. In the other, the desired mode, the voltages at B and C are of the same sign, being opposite to the voltage at A. By virtue of the fact that the gaps of the cavity of figure 2.1 face in opposite directions, however, the instantaneous voltages across the two gaps B and C are always in opposite sense. The phases at A, B and C for both modes were confirmed by measurement with a phase meter.

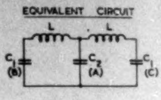


Figure 3.

The dimensions of the cavity (figure 2.1) were determined by the following considerations:—

(a) *Size of inner tube.* This was required to be large, to transmit a reasonable current of electrons; but small, to give maximum shunt resistance; 3/8″ O.D., 9/32″ I.D., was selected as the reasonable compromise.

(b) *Disc thickness and external wall diameter.* To give adequate conduction of the heat generated in the central baffles, the disc was chosen to be ¼″ thick. The gap size was desired small, so as to maximize acceleration over a small portion of the R.F. cycle, yet not so small as to cause field emission from the tube ends. A total gap size of ½″ (¼″ at each end of figure 2.1) and an external wall diameter of 2⅝″ approx. were decided upon. Figures 2.2 and 2.4 show the variation of resonant wave-length of the cavity 2.1 with the variation of various parameters in the cavity.

The assumption in the experiments was that the voltage at A, figure 2.1, would be small compared with the voltage developed at B or C, and that most of the available power would be employed in generating the voltage where it was required. This arose from consideration of the electrostatic capacities at A and B, or from a simple consideration of the relative diameter of disc and tube. That this elementary picture was incorrect was suggested by the following facts:—

(*a*) The fundamental and desired modes differed in frequency by 20%. This would require the ratio C_2/C_1 (figure 3) to be 22%.

(*b*) The resonant frequency sensitivity to variation of gap G (figure 2) is only one-half the resonant frequency sensitivity to variation of gap g.

(*c*) When a polythene rod was inserted so as to fill the tube and gap g, the frequency change was 6%; when a polythene annulus filled gap G, the frequency change was 3·8%.

It is difficult to give an accurate interpretation of these facts: but at least it would seem that the voltage at A is some 20–30% of the voltage at B or C.

The coupling to the cavity was first investigated by a single slot coupled to a simple re-entrant cavity (figure 4). It was found that the reflection coefficient was quite closely proportional to the slot length a over a wide range (30%): this would correspond to the coupling factor being proportional to a^n, where experiment showed $n = 2{\cdot}5$. The theory of coupling given by H. A. Bethe suggests that the coupling of a cavity to a guide by a slot is determined by the magnetic "polarizability" of the slot, which in the case of a narrow elliptical slot is given by

$$\frac{\pi}{3}\left[\frac{a^3}{\log_e(4a/b)-1}\right],$$

where b is the smaller semi-axis. In our case, neglecting the difference between rectangular and oval slots, this would give an effective value of n of 2.6. For the double slot and single cavity as shown in figure 4 the slot length required is 0·7″: for the 3-section cavity, the slot length was 0·95″. Generally speaking, the tolerance on the slot length was 0·02″.

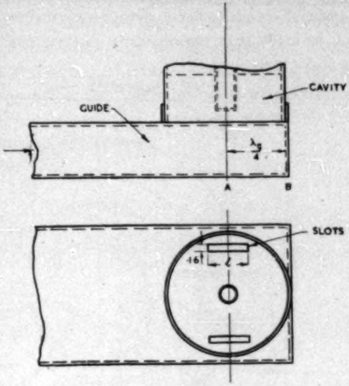


Figure 4.

The tuning of the 3-section cavity was accomplished by the plungers shown in figure 1 A; but as these were also in the region of some electric field, the degree of the tuning that could be achieved by them was only 0·5%. To ensure the correct initial frequency, the discs were at first made oversize, and then turned down until the required wave-length was reached (10·01 cm.). The tuning range was then 10·01–9·96, as required by the magnetron. The vacuum was maintained in the guide by a quartz window waxed to an inductive iris, the system as a whole giving negligible reflection. The 3-section cavity had, in addition to the desired mode (10 cm.) and the fundamental mode (12·1 cm.), two other modes: 10·4 and 11·4 cm. These, however, were all well outside the frequency range of the magnetron. The eventual Q of the cavity, when matched to the guide, was 4000: this reduced, after considerable operation (presumably due to sparking), to 3000.

§ 3. PHASING

The determination of cavity dimensions depends upon the voltage expected across the first few gaps. From the 25 cm. observations, it was considered that a total available peak power of 250 kw. into the cavity and effective shunt resistance of 0·5 MΩ. were reasonable minimum estimates. These figures gave a peak voltage of 300 kv. per gap, or 150 kv. across the first gap. The potential distribution across the gaps was approximated by a method due to R. D. Hill (1945), and step-by-step integration conducted across the gap. In this way, the incident and emergent phases were determined. The inter-gap distances were not critical, and were determined as $1\frac{1}{2}''$, $1\frac{5}{8}''$ and $1\frac{3}{4}''$. The electrons were quite well bunched between first and second gaps; thus, electrons entering the first gaps with phase between 0·4 and 1·2 radian entered the second gap with phase between 0·5 and 1·0 radians.

§4. HIGH-POWER WORK

The work on feeding the cavity with power followed the principles laid down by J. R. Pierce and developed in this Laboratory by B. Y. Mills. It can be described with reference to figure 5. The first step is to determine the *Rieke Diagram* of the magnetron, i.e. to plot on the Smith chart the contours of the frequency and power developed by the magnetron when looking at admittances corresponding to points on the chart. The admittance that the magnetron "looked at" was determined by replacing the magnetron by a low power C.W. coaxial standing-wave detector connected to a dummy output seal. This assumes, it is true, that the dummy output seal is identical with the magnetron output seal, which is highly unlikely. But we are mainly concerned with replacing the loads, with which the Rieke is determined, by the cavity, so that, since the method is essentially a substitution method, the difference between magnetron output and dummy output is negligible.

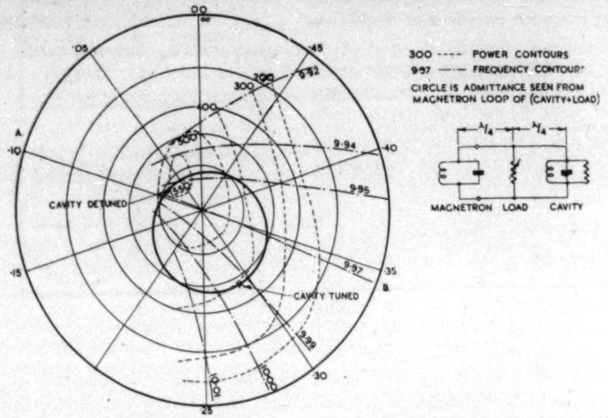

00
∞
·05
·45
A
·10
·40
·15
·35
B
·30
·25
200
300
400
500
550
9·92
9·94
9·95
9·97
9·99
10·00
10·01
CAVITY DETUNED
CAVITY TUNED
300 ---- POWER CONTOURS
9·97 —·— FREQUENCY CONTOURS
CIRCLE IS ADMITTANCE SEEN FROM MAGNETRON LOOP OF (CAVITY+LOAD)
λ/4
λ/4
MAGNETRON
LOAD
CAVITY

Figure 5. Rieke diagram. cv75. 40A. 2300 gauss.

Figure 5, then, is the Rieke diagram referred to the magnetron output loop. The important point here is the region of the "Frequency sink" (approximately marked "Cavity detuned"); here the frequency contours tend to cross over, and the frequency becomes indeterminate. It was necessary to ascertain this point with some care, for it is at this point that the magnetron circuit (figure 5, inset) can be regarded as being located. Suppose now we have a cavity which when tuned is matched to the magnetron output, and we place the cavity at an even number of half wave-lengths from the magnetron. Then the condition for stable operation is that the admittance change with frequency shall be opposite in direction to the change of the frequency contours which it crosses. The point of match will be a stable point of operation. But the admittance circle of the cavity, with varying frequency, touches the edge of the diagram at A: and by virtue of the crossing contours at the frequency sink, there are also other stable points of operation at points far off the resonant frequency of the cavity, at which the magnetron is developing low power. It is on these modes that the magnetron prefers to operate, if the cavity is half a wave-length from the magnetron. If it is a quarter wave-length, then the point of match is definitely unstable; the magnetron will operate, again, only on points near the detuned point B on the diagram.

The only choice, therefore, is to make the magnetron operate with the cavity half a wave-length away, and yet move the detuned point inside the frequency sink. This is done by inserting a resistive (water) load at the quarter wave point. The resulting admittance circle of (cavity + load) obtained by varying the cavity resonance, is shown by the thick black line of figure 5. As a consequence of the inclusion of the load, a fraction of the power is absorbed by the load, and the magnetron works at a point appreciably off the point of maximum power development; so that, with a 500-kw. magnetron, a maximum of 300 kw. only was eventually available for acceleration of electrons.

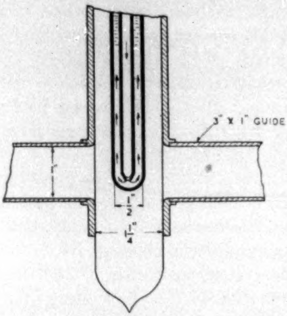


Figure 6 (*a*).

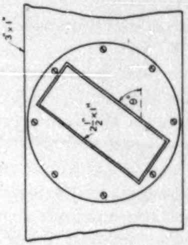


Figure 6 (*b*).

For purposes of experiment, it was desirable to have a high-power load which behaved as a pure variable conductance. Two such loads were developed and are shown in figure 6. The first is simply a water tube which can be pushed in and out of the centre of the wide face of the guide, the coupling varying with the depth of penetration. The tube was $\frac{1}{2}''$ diameter, and was carried in a jacket of Freon, to minimize sparking. It behaved very nearly as a pure conductance over the 0·25 to $1{\cdot}25Y_0$ and, in view of the ease of manipulation, was used throughout the experiments. The other, possibly a better load for higher powers, was essentially a T-junction on the broad side of the $3'' \times 1''$ guide, the side arm being $2\frac{1}{2}'' \times 1''$ guide terminated with a matched water load. When in the normal position, as an E-plane stub, this load was very nearly Z_0 in series with the arterial $3'' \times 1''$ guide: when at right angles to this, the series impedance was negligible. Intermediate positions showed series impedances closely proportional to $\sin^2\theta$, as one would expect.

Observations during operation showed that the ratio of the power in the water load to that in the cavity was considerably higher than that expected from the C.W. measurement.

It was decided that this was due to three effects:—

1. Build-up in the cavity reduced the mean power developed in it.

2. After the magnetron pulse is over, some of the decaying oscillation in the cavity is fed back through the guide to the water load.

3. The main point is that at the beginning of the pulse the current has not built up in the cavity, which is thus mismatched to the guide. The admittance presented to the magnetron is thus effectively the detuned admittance, where the power developed is 500 kw., all of which is absorbed by the water load. The mean power in the load is, therefore, much higher than the instantaneous power at the end of the pulse, which is, of course, the power relevant for acceleration purposes.

§ 5. ACCELERATION EXPERIMENTS

The electrons to be accelerated were provided by an electron gun, which was initially operated at 12 kv. DC. After passing through the cavity and guide (figure 1 A) they passed into a magnetic analyser. Comparison between the magnetic field and applied voltage at 12 kv. showed agreement within 3%, which is probably within the limits of accuracy of the various meters concerned. A feature at both high (800 kv.) and low voltage observations was a "tail", about 1% of the peak current, extending to some 20–30% of the peak field on either side of the peak; it was presumed to be due to reflection at the edges of the defining slit of the analyser.

During the acceleration experiments, the 25 kv. magnetron pulse was applied to the gun, and the current to the slit was initially picked up by the pulse amplifier. The voltage of 800–850 kv. was not materially changed by adjustment of many of the available variables; the peak was quite sharp, although on the low-voltage side there was a considerable number of electrons (order of 10^{-8} amp. mean) provided by emission of one kind or another in the accelerating gaps. When the current was maximized, 3μA. mean was available at the analyser slit, the voltage dropping to some 800 kv. The duty cycle was 2800, so that some 8 mA. peak was reaching the slit; and as the cavity does not reach peak voltage till towards the end of the pulse, and as acceleration takes place only over a fraction of the R.F. cycle, the instantaneous current is an appreciable fraction of 1 ampere. Considering also that the analyser slit was $\frac{5}{8}'' \times \frac{1}{8}''$ located 18″ from the orifice of the cavity, one concludes that either the focusing of cavity and analyser was good, or that the total peak current was considerable.

When the cavity was examined, it was found that considerable sparking had taken place between discs and cavity wall, particularly at the one next to the guide: as the cavity was effectively being evacuated through this annulus, presumably the pressure here at times could become appreciable. There was some evidence of anti-symmetric voltages at the disc, possibly generated by the tuning plungers, which were projecting considerably into the guide.

The figure of 1100 kv. mentioned in the abstract was arrived at as follows:— The observed maximum power in the cavity was 300 kw. peak, and an estimated shunt resistance of 1 MΩ (compared with some measured 25 cm. values) seemed attainable. These figures give, for each stage, an R.M.S. voltage of 310 kv. or a peak voltage of 450 kv.: which for 3 gaps gives a maximum possible of 1350 kv.

Of this, the peak current, as calculation shows, is 90% of the maximum possible: and the remaining deficit between 850 kv, and 1220 kv. is presumably to be ascribed either to sparking at axial or peripheral gaps, or to the appreciable proportion of the available voltage being taken up at the peripheral gap. Unfortunately circumstances prevented an adequate analysis of these alternatives being made.

ACKNOWLEDGMENT

The work was carried out as part of the programme of the Radiophysics Division of C.S.I.R., Australia.

REFERENCES

Bethe, H. A., *M.I.T. Report*, **43**, 22.

Bowen, Pulley and Gooden, 1946. *Nature, Lond.*, **157**, 840.

Hill, R. D., 1945. *J. Sci. Instrum.*, **22**, 221.

Lawrence, E. O. and Sloan, D. H., 1931. *Phys. Rev.*, **38**, 2021.

Pierce, J. R. B.T.L. MM 43/140/19.

Sloan, D. H., 1935. *Phys. Rev.*, **47**, 62.

THERMODYNAMIC RELATIONS FOR TWO PHASES CONTAINING TWO COMPONENTS IN EQUILIBRIUM UNDER GENERALIZED STRESS

By C. GURNEY,

Royal Aircraft Establishment, Farnborough

MS. received 7 *December* 1946

ABSTRACT. The application of thermodynamics to cases of other than hydrostatic pressure is important in connection with the swelling and flow and fracture of solids under generalized stress. In the present paper the methods of Gibbs are applied to the case of two phases containing the same two component substances in equilibrium with each other.

The problem is first considered in its most general form, each phase being under generalized stress and each containing each component. The more particular problem in which one of the components is absent from one of the phases is then considered, and the particular case in which one of the phases is fluid and, therefore, able to withstand only hydrostatic pressure, is dealt with in some detail. The cases of a two-component fluid phase in equilibrium with a one-component solid phase and a one-component fluid phase in equilibrium with a two-component solid phase are treated together. These cases correspond respectively to what are often called solution and swelling, although there is no logical reason for this nomenclature. The derivatives of pressure on the fluid phase for changes of temperature and changes of each of the components of generalized stress on the solid phase are given. When suitably interpreted, the same formulae apply to both solution and swelling. Formulae for entropy changes with stress and temperature are also given, and the use of other independent variables such as strain, force, and displacement instead of stress is discussed.

§ 1. INTRODUCTION

The most usual stress system considered in thermodynamics is hydrostatic pressure. Solids can withstand generalized stress, and increasing interest is being shown in the thermodynamics of stress systems involving other than hydrostatic pressure. This subject finds applications in the swelling of substances such as wood and plastics by water, and in the flow of rocks, where in some cases the flow is attributed to solution of highly stressed parts of the rock and deposition of the dissolved material at stress-free places; and it is likely to be important in connection with the fracture of brittle materials which are subject to attack by the surrounding medium. Here the increasing severity of the attack with increasing stress is thought to cause the highly stressed material at the ends of cracks to be preferentially attacked, so that the cracks gradually spread and cause delayed fracture. The subject should also be of importance in metallurgy, where internal stresses due to work hardening or anisotropic thermal expansion may decrease the stability of the structure, and lead to weakening or the development of cracks; and phase transitions due to stress may lead to flow and failure.

The application of thermodynamics to stress systems other than hydrostatic pressure has already been discussed by Gibbs (1876) in his original paper on the equilibrium of heterogeneous substances. In the present paper the methods developed by Gibbs are applied to the case of equilibrium between two phases containing the same two components. The subject is first discussed in its most general aspect, each phase containing each component and each subjected to its own system of generalized stress. On account of the need to satisfy the equality of the chemical potential of each component in both phases, it is not possible to vary only two independent variables at a time, and true partial derivatives of the variables cannot be obtained. It is therefore not possible to arrive at results of much generality. If, however, one of the components is absent from one of the phases, it is only necessary to satisfy the conditions of equality of the chemical potentials of the component common to both phases. For this case, true partial derivatives can be obtained. Two cases of this sort are of practical importance:—(1) A solid two-component phase in equilibrium with a fluid one-component phase: the absorption of water by wood is an example of this sort. (2) The other case is that of a one-component solid phase in equilibrium with a two-component fluid phase: saturated solutions of many salts provide examples of this. For these two cases the partial derivatives of the pressure on the fluid phase with respect to the temperature and with respect to any one of the components of generalized stress acting on the solid phase have been computed. It is of some interest that when suitably interpreted the same formulae apply to both of these cases.

§ 2. THERMODYNAMIC RELATIONS FOR SYSTEM UNDER GENERALIZED STRESS

The stress system acting on a body enters thermodynamics via the work which forces do in moving their points of application. Two of the four thermodynamic energy functions contain "work done" explicitly. These are the energy E and the Helmholtz free energy F. The energy also contains entropy changes explicitly, and is therefore useful when considering rapid changes which may be assumed to take place adiabatically, for such changes take place at constant entropy. We are here more concerned with slow changes which take place isothermally, and we therefore choose F as our energy function because it contains temperature change explicitly. For a phase containing two components, the most general change in F is given by

$$dF = -SdT + dW + \mu_1 dn_1 + \mu_2 dn_2. \qquad \text{......(1)}$$

Here S is entropy, T is temperature, W is work, μ is chemical potential, n is quantity of component. The two components are designated by subscripts 1 and 2.

For hydrostatic pressure the change in work done, dW, is equal to $-pdV$, where V is volume and p is pressure. If we wish to change the independent variable to p, we write $dW = -p\frac{\partial V}{\partial p}dp$. For work done under generalized stress we have similar choice of variables. We may choose the six components of force (three direct forces and three shear forces), and with these it is convenient to choose six components of displacement as associated variables. The components are conveniently distinguished by two numerical suffixes, the first suffix indicating the direction of the normal to the plane on which the force acts and the second suffix the direction of the force. Thus if P and dx are forces and displacements respectively, P_{11}, P_{22}, P_{33}, dx_{11}, dx_{22}, dx_{33} are direct forces and extensional displacements, while P_{12}, P_{23}, P_{31}, dx_{12}, dx_{23}, dx_{31} are shear forces and shear displacements. It is convenient to adopt the shorthand notations P_{ij} and dx_{ij} for force and displacement, where i and j are understood to stand for numbers between 1 and 3. Tension is taken as positive. The work done for a general small change in displacements is then

$$dW = \Sigma_{ij}P_{ij}dx_{ij}. \qquad \ldots\ldots(2)$$

If it is desired to have the P_{ij} as the independent variables we may write for constant temperature and quantities of each component

$$dx_{ij} = \Sigma_{kl}\frac{\partial x_{ij}}{\partial P_{kl}}dP_{kl} \qquad \ldots\ldots(3)$$

and

$$dW = \Sigma_{ij}P_{ij}\Sigma_{kl}\frac{\partial x_{ij}}{\partial P_{kl}}dP_{kl}. \qquad \ldots\ldots(4)$$

Instead of force and displacement it is more usual to use stress and strain as variables. These may be denoted by X_{ij} and e_{ij}. In terms of these variables, change in work done becomes

$$dW = V\Sigma_{ij}X_{ij}de_{ij}, \qquad \ldots\ldots(5)$$

where V is the volume. If we wish to have stress as independent variable, the change in W becomes

$$dW = V\Sigma_{ij}X_{ij}\Sigma_{kl}\frac{\partial e_{ij}}{\partial X_{kl}}dX_{kl}. \qquad \ldots\ldots(6)$$

In the theory of elasticity, strain is the fractional change of length due to stress, but here we use strain to denote proportional change in length due to any cause. It will therefore include change in length due to change in temperature, due to change in composition and due to change in quantity of the phase. The most general expression for dW is, therefore,

$$dW = V\Sigma_{ij}X_{ij}\left(\Sigma_{kl}\frac{\partial e_{ij}}{\partial X_{kl}}dX_{kl} + \frac{\partial e_{ij}}{\partial T}dT + \frac{\partial e_{ij}}{\partial n_1}dn_1 + \frac{\partial e_{ij}}{\partial n_2}dn_2\right). \qquad \ldots\ldots(7)$$

Substituting this value in expression (1) gives

$$dF = \left(-S + V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial T}\right)dT + V\Sigma X_{ij}\Sigma\frac{\partial e_{ij}}{\partial X_{kl}}dX_{kl} + \left(\mu_1 + V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n_1}\right)dn_1$$
$$+ \left(\mu_2 + V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n_2}\right)dn_2. \qquad \ldots\ldots(8)$$

For equilibrium between two phases each containing two components, the following relations must hold (Gibbs, 1876):

$$T^{\alpha}=T^{\beta}; \quad \mu_1^{\alpha}=\mu_1^{\beta}; \quad \mu_2^{\alpha}=\mu_2^{\beta}, \qquad \ldots\ldots(9)$$

where the superscripts α and β designate the phases and the suffixes 1 and 2 designate the components. After any changes in the variables of the system these relations must continue to be valid. Suppose each of the phases is under generalized stress and that we vary one stress denoted by X_{kl} of the phase α and another stress denoted by X_{mn} of the phase β. Then we have

$$\frac{\partial\mu_1}{\partial X_{kl}}\,dX_{k}^{\alpha}=\frac{\partial\mu_1}{\partial X_{mn}}\,dX_{mn}^{\beta}, \qquad \ldots\ldots(10)$$

$$\frac{\partial\mu_2}{\partial X_{kl}}\,dX_{kl}^{\alpha}=\frac{\partial\mu_2}{\partial X_{mn}}\,dX_{mn}^{\beta}. \qquad \ldots\ldots(11)$$

Obviously in general these equations cannot simultaneously be satisfied, and we are therefore not free to vary only one variable of each phase at a time. Some other variable of one of the phases, such as another stress or temperature or quantity of the components, must be varied. The partial change of a variable of one phase cannot therefore be expressed with respect to a variable of the other phase, all other variables remaining constant. In the general case, therefore, results of sufficient generality to justify their inclusion in this paper cannot be obtained.

If, however, one of the phases contains only one component, say component 1, it is only necessary to ensure equality of the chemical potentials of that component. This is a case of common occurrence in practice. It includes, for example, the equilibrium of wood containing absorbed moisture with water, and the equilibrium of many salts with their saturated solutions; and it includes many examples found in metallurgy. We then have for the partial variation in stresses on the two phases

$$\frac{\partial X_{kl}^{\alpha}}{\partial X_{mn}^{\beta}}=\frac{\dfrac{\partial\mu_1^{\beta}}{\partial X_{mn}}}{\dfrac{\partial\mu_1^{\alpha}}{\partial X_{kl}}}. \qquad \ldots\ldots(12)$$

We may evaluate this expression in terms of other variables of the system by using the fact that dF is a total differential and, therefore, that the order of differentiation is of no consequence. Differentiating the third coefficient of expression (8) with respect to stress and the second with respect to quantity of component 1 gives

$$\frac{\partial\mu_1}{\partial X_{kl}}=\frac{\partial}{\partial n_1}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{kl}}\right)-\frac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n_1}\right). \qquad \ldots\ldots(13)$$

Expression (12) therefore becomes

$$\frac{\partial X_{kl}^{\alpha}}{\partial X_{mn}^{\beta}}=\frac{\left[\dfrac{\partial}{\partial n_1}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial X_{mn}}\right)-\dfrac{\partial}{\partial X_{mn}}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial n_1}\right)\right]^{\beta}}{\left[\dfrac{\partial}{\partial n_1}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial X_{kl}}\right)-\dfrac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial n_1}\right)\right]^{\alpha}}. \qquad \ldots\ldots(14)$$

Instead of varying the stress on both phases we may obtain equilibrium by varying two stress components on one phase, so that the resultant change in μ is zero. We would then have an expression similar to expression (14) with α substituted for β and the sign changed.

If, when the stress on the α phase is varied, other stresses are maintained constant and the temperature is varied to maintain equilibrium, we have

$$\left[\frac{\partial\mu_1}{\partial X_{kl}}dX_{kl}+\frac{\partial\mu_1}{\partial T}dT\right]^{\alpha}=\left[\frac{\partial\mu_1}{\partial T}dT\right]^{\beta}. \qquad \ldots\ldots(15)$$

Thus

$$\frac{\partial X_{kl}}{\partial T}=\frac{\dfrac{\partial\mu_1^{\beta}}{\partial T}-\dfrac{\partial\mu_1^{\alpha}}{\partial T}}{\dfrac{\partial\mu_1^{\alpha}}{\partial X_{kl}}}. \qquad \ldots\ldots(16)$$

By differentiating the third coefficient of expression (8) with respect to T and the first with respect to n and equating, and using the previously obtained expression for $\dfrac{\partial\mu_1}{\partial X_{kl}}$, equation (16) becomes

$$\frac{\partial X_{kl}}{\partial T}=\frac{\Delta\left[-\dfrac{\partial}{\partial T}V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial n_1}-\dfrac{\partial S}{\partial n_1}+\dfrac{\partial}{\partial n_1}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial T}\right)\right]}{\left[\dfrac{\partial}{\partial n_1}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial X_{kl}}\right)-\dfrac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\dfrac{\partial e_{ij}}{\partial n_1}\right)\right]^{\alpha}}. \qquad \ldots\ldots(17)$$

Here the symbol Δ indicates the difference between corresponding quantities for the phases β and α. $T\Delta\dfrac{\partial S}{\partial n_1}$ is the latent heat of isothermal change of unit quantity of component 1 from a large quantity of the phase α to phase β.

When the stress on one phase is varied, we may also obtain equilibrium by varying the composition of either phase, the other variables being kept constant. Thus when the composition of the other phase is varied

$$\left[\frac{\partial\mu_1}{\partial X_{kl}}\right]^{\alpha}dX_{kl}=\left[\frac{\partial\mu_1}{\partial n_1}dn_1\right]^{\beta}, \qquad \ldots\ldots(18)$$

giving

$$\frac{[\partial X_{kl}]^{\alpha}}{[\partial n_1]^{\beta}}=\frac{\left[\dfrac{\partial\mu}{\partial n_1}\right]^{\beta}}{\left[\dfrac{\partial\mu_1}{\partial X_{kl}}\right]^{\alpha}}. \qquad \ldots\ldots(19)$$

This expression is not as useful as those previously obtained, as $\dfrac{\partial\mu}{\partial n_1}$ cannot be equated to expressions containing derivatives of strain.

Having thus obtained expressions for the partial derivative of stress with respect to other components of stress, temperature and composition, we may obtain expressions for the partial derivatives of any of the independent variables in terms of any other independent variables. For example, we may obtain the partial variation of composition with temperature from the expression

$$\frac{\partial n_1}{\partial T}=-\frac{\dfrac{\partial X_{kl}}{\partial T}}{\dfrac{\partial X_{kl}}{\partial n_1}}. \qquad \ldots\ldots(20)$$

So far we have considered both phases being acted upon by generalized stress. If one phase is a fluid, and can therefore only permanently withstand hydrostatic pressure we could obtain the partial derivatives of pressure with respect to the other variables by varying X_{11}, X_{22}, and X_{33} by equal amounts simultaneously. Thus

$$\frac{\partial\mu}{\partial X_{11}}dX_{11}+\frac{\partial\mu}{\partial X_{22}}dX_{22}+\frac{\partial\mu}{\partial X_{33}}dX_{33}=\frac{\partial\mu}{\partial p}dp \qquad \ldots\ldots(21)$$

if

$$X_{11}=X_{22}=X_{33}=p \quad \text{and} \quad dX_{11}=dX_{22}=dX_{33}=dp.$$

Alternatively we may write for the fluid phase

$$dF=\left[\left(-S-p\frac{\partial V}{\partial T}\right)dT-p\frac{\partial V}{\partial p}dp+\left(\mu_1-p\frac{\partial V}{\partial n_1}\right)dn_1+\left(\mu_2-p\frac{\partial V}{\partial n_2}\right)dn_2\right]^{\mathrm{F}} \qquad \ldots\ldots(22)$$

and evaluate $\partial\mu/\partial p$ from cross differentiating appropriate terms of expression (22).

Two cases of a solid phase in equilibrium with a fluid phase are of particular interest. The first is a two-component solid phase in equilibrium with a one-component fluid phase: this is a common case in swelling phenomena. The other case is a one-component solid phase in equilibrium with a two-component fluid phase: this corresponds to equilibrium between a pure solid substance and its saturated solution. As changes in chemical potential and quantity only occur for the substance which is common to both phases, we may drop suffixes and write

$$\mu^{\mathrm{S}}=\mu^{\mathrm{F}}, \qquad \ldots\ldots(23)$$

where μ is the chemical potential of the component common to both phases; the superscripts S and F indicate solid and fluid phases. In the same way $\partial\mu/\partial n$ must be understood to mean the derivative of the chemical potential of the component which is common to both phases with respect to the quantity of this component. With this notation, the same general formulae apply to both the particular cases mentioned above. In the next section we discuss these cases in more detail.

§ 3. DERIVATIVES OF PRESSURE FOR SOLID-FLUID EQUILIBRIUM

In this section we discuss a two-component solid phase in equilibrium with a one-component fluid phase, and a one-component solid phase in equilibrium with a two-component fluid phase. The same formulae apply. For brevity we refer to the former as a case of swelling and the latter as solution, although logically either of these terms might be applied to either of the cases considered.

3.1. *Variation of pressure with stress.*

By the method of § 2 we readily obtain for the variation in pressure on the fluid phase with stress on the solid phase

$$\frac{\partial p}{\partial X_{kl}}=\frac{\left[\frac{\partial}{\partial n}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{kl}}\right)-\frac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n}\right)\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(24)$$

This may be rewritten

$$\frac{\partial p}{\partial X_{kl}} = \frac{\left[\frac{\partial V}{\partial n}\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{kl}} - \frac{\partial V}{\partial X_{kl}}\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n} - V\frac{\partial e_{kl}}{\partial n}\right]^{S}}{\left[\frac{\partial V}{\partial n}\right]^{F}}. \qquad \ldots\ldots(25)$$

The terms such as $\frac{\partial V}{\partial n}$, $\frac{\partial e}{\partial X}$ etc. in this and subsequent equations are all functions of the independent variables.

It is of interest to discuss the physical significance of the terms in equation(25). In the case of swelling, $\frac{\partial V^{S}}{\partial n}$ is the volume swelling of the solid per unit mass of absorbed fluid. $\frac{\partial e_{ij}}{\partial X_{kl}}$ are the reciprocals of the elastic constants of the solid, such as $\frac{1}{E}$ or $-\frac{\nu}{E}$. $\frac{\partial V}{\partial X_{kl}}$ is the volume change due to change in the particular stress under consideration. $\frac{\partial e_{ij}}{\partial n}$ are the changes in strain due to swelling, and $\frac{\partial e_{kl}}{\partial n}$ are the changes (due to swelling) in the strain corresponding to the particular stress which is varied. $\left[\frac{\partial V}{\partial n}\right]^{F}$ is the specific volume of the fluid.

In the case of solution some of the terms have very different significance. In this case $\frac{\partial V^{S}}{\partial n}$ is the specific volume of the solid, and $\frac{\partial e_{ij}}{\partial n}$ and $\frac{\partial e_{kl}}{\partial n}$ are the changes in proportional dimensions due to the solid being dissolved away. $\left[\frac{\partial V}{\partial n}\right]^{F}$ is the volume swelling of the fluid per unit mass of dissolved solid.

It is of interest to consider some simple cases:—

Case 1. *Simple direct stress*

$$X_{22} = X_{33} = X_{12} = X_{23} = X_{31} = 0.$$

$$\frac{\partial p}{\partial X_{11}} = \frac{\left[\frac{\partial V}{\partial n}X_{11}\frac{\partial e_{11}}{\partial X_{11}} - \frac{\partial V}{\partial X_{11}}X_{11}\frac{\partial e_{11}}{\partial n} - V\frac{\partial e_{11}}{\partial n}\right]^{S}}{\left[\frac{\partial V}{\partial n}\right]^{F}}. \qquad \ldots\ldots(26)$$

In the case of swelling of an isotropic material with shear modulus G and linear strain e and specific volume of pure fluid $\overline{V}$,

$$\frac{\partial p}{\partial X_{11}} = \frac{\left[V\frac{\partial e}{\partial n}\left(\frac{X_{11}}{G} - 1\right)\right]^{S}}{[\overline{V}]^{F}}. \qquad \ldots\ldots(27)$$

In the case of solution off the face on which X_{11} acts,

$$\frac{\partial p}{\partial X_{11}} = \frac{\left[\overline{V}\left(\frac{2\nu X_{11}}{E} - 1\right)\right]^{S}}{\left[\frac{\partial V}{\partial n}\right]^{F}}, \qquad \ldots\ldots(28)$$

where $\overline{V}$ is the specific volume of pure solid. If solution takes place off a stress-free face then $\frac{\partial e_{11}}{\partial n}=0$ and expression (26) becomes

$$\frac{\partial p}{\partial X_{11}}=\frac{\left[\overline{V}\frac{X_{11}}{E}\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(29)$$

For stressses small compared with the elastic modulus, expression (28) usually has the biggest value, expression (27) the next biggest, and expression (29) is smallest. The latter is zero at zero stress. At low stresses, increase in compression produces effects of opposite sign to increase in tension for the cases represented by equations (27) and (28). The effect represented by equation (29) is independent of the sign of the stress.

Case 2. *Simple shear stress*

$$X_{11}=X_{22}=X_{33}=X_{23}=X_{31}=0.$$

Equation (25) gives

$$\frac{\partial p}{\partial X_{12}}=\frac{\left[\frac{\partial V}{\partial n}X_{12}\frac{\partial e_{12}}{\partial X_{12}}-\frac{\partial V}{\partial X_{12}}X_{12}\frac{\partial e_{12}}{\partial n}-V\frac{\partial e_{12}}{\partial n}\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}} \qquad \ldots\ldots(30)$$

In the case of swelling of an isotropic material, the value of $\frac{\partial e_{12}}{\partial n}$ depends on the rate of change of shear modulus with moisture content and equals $-\frac{X_{12}}{G}\frac{\partial G}{\partial n}$. The middle term is then negligible, and if e is the linear swelling

$$\frac{\partial p}{\partial X_{12}}=\frac{\left[\frac{V}{G}X_{12}\left(3\frac{\partial e}{\partial n}+\frac{1}{G}\frac{\partial G}{\partial n}\right)\right]^{\mathrm{S}}}{[\overline{V}]^{\mathrm{F}}} \qquad \ldots\ldots(31)$$

In the case of solution off any face, $\frac{\partial e_{12}}{\partial n}$ is zero and (30) becomes

$$\frac{\partial p}{\partial X_{12}}=\frac{\left[\overline{V}\frac{X_{12}}{G}\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(32)$$

Case 3. *Three unequal principal stresses*

$$X_{12}=X_{23}=X_{31}=0.$$

This is only discussed for the case of swelling. Solubility would be different on each of the three pairs of opposite faces of a cube, and to obtain equilibrium the liquid in contact with each opposite pair would have to be isolated from that in contact with other pairs. Solubility for this case is not, therefore, further discussed. For swelling

$$\frac{\partial p}{\partial X_{11}}=\frac{V^{S}\left[X_{11}\left\{-\frac{\partial e_{11}}{\partial n}\left(\frac{\partial e_{22}}{\partial X_{11}}+\frac{\partial e_{33}}{\partial X_{11}}\right)+\frac{\partial e_{11}}{\partial X_{11}}\left(\frac{\partial e_{22}}{\partial n}+\frac{\partial e_{33}}{\partial n}\right)\right\}+\begin{matrix}\text{similar terms}\\ \text{in } X_{22} \text{ and } X_{33}\end{matrix}-\frac{\partial e_{11}}{\partial n}\right]^{S}}{\left[\frac{\partial V}{\partial n}\right]^{F}}. \qquad \ldots\ldots(33)$$

For an isotropic material

$$\frac{\partial p}{\partial X_{11}}=\frac{\left[V\frac{\partial e}{\partial n}\left\{\frac{2X_{11}-\overline{X_{22}+X_{33}}}{2G}-1\right\}\right]^{S}}{\left[\frac{\partial V}{\partial n}\right]^{F}}. \qquad \ldots\ldots(34)$$

The total change due to changes in the three stresses is

$$dp=\frac{\partial p}{\partial X_{11}}dX_{11}+\frac{\partial p}{\partial X_{22}}dX_{22}+\frac{\partial p}{\partial X_{33}}dX_{33}. \qquad \ldots\ldots(35)$$

If in equation (33) we neglect all the terms in the numerator except $-\frac{V\partial e_{11}}{\partial n}$, equation (35) becomes

$$\left[\frac{\partial V}{\partial n}\right]^{F}dp=\left[-V\left\{\frac{\partial e_{11}}{\partial n}dX_{11}+\frac{\partial e_{22}}{\partial n}dX_{22}+\frac{\partial e_{33}}{\partial n}dX_{33}\right\}\right]^{S}. \qquad \ldots\ldots(36)$$

This is Barkas's equation (12) in his 1945 paper. It applies so long as the stresses are small compared with the shear modulus or to an isotropic material if the initial state is one of hydrostatic pressure. This conclusion has been reached by Warburton (1946) by a different analysis.

3.2. *Variation of pressure with temperature*

By the method of § 2 we obtain for the variation of the pressure on the fluid phase with the variation of the common temperature of both phases when the solid phase is under generalized stress,

$$\frac{\partial p}{\partial T}=\frac{\left[\frac{\partial V}{\partial n}\Sigma X_{ij}\frac{\partial e_{ij}}{\partial T}-\frac{\partial V}{\partial T}\Sigma X_{ij}\frac{\partial e_{ij}}{\partial n}\right]^{S}+\frac{\partial}{\partial n}\frac{L}{T}}{\left[\frac{\partial V}{\partial n}\right]^{F}}. \qquad \ldots\ldots(37)$$

$\frac{\partial L}{\partial n}$ is the latent heat of transfer of unit quantity of the component common to both phases from a large quantity of the solid phase to a large quantity of the fluid phase.

Case 1. *Simple tension*

$$X_{22}=X_{33}=X_{12}=X_{23}=X_{31}=0.$$

$$\frac{\partial p}{\partial T}=\frac{\left[\frac{\partial V}{\partial n}X_{11}\frac{\partial e_{11}}{\partial T}-\frac{\partial V}{\partial T}X_{11}\frac{\partial e_{11}}{\partial n}\right]^{S}+\frac{\partial}{\partial n}\frac{L}{T}}{\left[\frac{\partial V}{\partial n}\right]^{F}}; \qquad \ldots\ldots(38)$$

for an isotropic material and in the case of swelling

$$\frac{\partial p}{\partial T}=\frac{\left[\frac{1}{T}\frac{\partial L}{\partial n}\right]}{[V]^{\mathrm{F}}}; \qquad \ldots\ldots(39)$$

for an isotropic material and in the case of solution off the stressed faces

$$\frac{\partial p}{\partial T}=\frac{\left[-2\bar{V}\frac{\partial e}{\partial T}X_{11}\right]^{\mathrm{S}}+\frac{1}{T}\frac{\partial L}{\partial n}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}; \qquad \ldots\ldots(40)$$

for an isotropic material and in the case of solution off the stress-free faces,

$$\frac{\partial p}{\partial T}=\frac{\left[\bar{V}\frac{\partial e}{\partial T}X_{11}\right]^{\mathrm{S}}+\frac{1}{T}\frac{\partial L}{\partial n}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(41)$$

Case 2. *Simple shear*

$$X_{11}=X_{22}=X_{33}=X_{23}=X_{31}=0.$$

$$\frac{\partial p}{\partial T}=\frac{\left[\frac{\partial V}{\partial n}X_{12}\frac{\partial e_{12}}{\partial T}-\frac{\partial V}{\partial T}X_{12}\frac{\partial e_{12}}{\partial n}\right]^{\mathrm{S}}+\frac{1}{T}\frac{\partial L}{\partial n}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(42)$$

If we neglect the variation in shear modulus with temperature and composition, equation (42) becomes for an isotropic material and in the case of swelling

$$\frac{\partial p}{\partial T}=\frac{\frac{1}{T}\frac{\partial L}{\partial n}}{[V]^{\mathrm{F}}}, \qquad \ldots\ldots(43)$$

and in the case of solution off any face

$$\frac{\partial p}{\partial T}=\frac{\frac{1}{T}\frac{\partial L}{\partial n}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(44)$$

The major effect of stress system on the value of $\frac{\partial p}{\partial T}$ is, therefore, its effect on the latent heat.

3.3. *Variation of pressure with composition*

Here it will be necessary to treat swelling and solution separately.

For swelling we have

$$\frac{\partial \mu^{\mathrm{S}}}{\partial n}\,dn=\frac{\partial \mu^{\mathrm{F}}}{\partial p}\,dp, \qquad \ldots\ldots(45)$$

$$\frac{\partial p}{\partial n}=\frac{\left[\frac{\partial \mu}{\partial n}\right]^{\mathrm{S}}}{[V]^{\mathrm{F}}}, \qquad \ldots\ldots(46)$$

where $\frac{\partial p}{\partial n}$ is the change in pressure on the fluid for an increase in the quantity of the common component absorbed by the solid.

For solution we have

$$\frac{\partial \mu}{\partial n} dn^{\mathrm{F}} + \frac{\partial \mu^{\mathrm{F}}}{\partial p} dp = 0, \qquad \text{......(47)}$$

$$\frac{\partial p}{\partial n} = - \frac{\left[\frac{\partial \mu}{\partial n}\right]^{\mathrm{F}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}} \qquad \text{......(48)}$$

giving the change in pressure on the fluid with increase in the quantity of the solid dissolved in the fluid. These expressions are not very useful in evaluating $\frac{\partial p}{\partial n}$, as $\frac{\partial \mu}{\partial n}$ is not a quantity directly measured experimentally. They are useful, however, in connections with evaluation of other derivatives.

3.4. *Miscellaneous*

In the expressions so far obtained, the pressure on the fluid has been assumed not to act on the solid. If it does act on the solid, then in the case of solution it can only be permitted to act on two opposite faces of the solid as the equilibrium pressure is different for the three opposite pairs of faces. If the normal to the faces on which it acts is denoted by m, which may take values 1, 2 or 3, then the pressure exerts tensile stress on these faces equal to $X_{mm} = -p$. The value of $\frac{\partial p}{\partial X_{kl}}$ (X_{mm} varying so as to equal $-p$) is equal to $\frac{\partial p}{\partial X_{kl}}$ (X_{mm} constant) divided by the factor $\left(1 + \frac{\partial p}{\partial X_{mm}}\right)$. In the case of swelling, the pressure may be allowed to act on all three faces and the corresponding dividing factor becomes $\left(1 + \sum_{1}^{3} \frac{\partial p}{\partial X_{ii}}\right)$. The above corrections apply for $mm \neq kl$. For $mm = kl$, the expressions of (3.1)–(3.3) apply without correction, it being understood that X_{mm} is the resultant of the applied normal stress and pressure of the fluid. The same dividing factors should be used to correct $\frac{\partial p}{\partial T}$.

The derivatives of the pressure on the fluid with respect to the other variables of the system—stress on solid, temperature, and composition of phases—have now been given. These derivatives may be used to calculate other quantities. For example, the variation in composition with temperature may be computed as follows:

$$dp = \frac{\partial p}{\partial X_{kl}} dX_{kl} + \frac{\partial p}{\partial T} dT + \frac{\partial p}{\partial n} dn. \qquad \text{......(49)}$$

For constant p and X_{kl} these are

$$\frac{\partial n}{\partial T} = \frac{-\frac{\partial p}{\partial T}}{\frac{\partial p}{\partial n}}. \qquad \text{......(50)}$$

For swelling, using equations (39) and (46),

$$\frac{\partial n}{\partial T}=\frac{-\dfrac{1}{T}\dfrac{\partial L}{\partial n}}{\left[\dfrac{\partial \mu}{\partial n}\right]^{\mathrm{S}}}. \qquad \ldots\ldots(51)$$

For solution at low stress, using equations (40) and (48) and neglecting the first term in the numerator of (40),

$$\frac{\partial n}{\partial T}=\frac{\dfrac{1}{T}\dfrac{\partial L}{\partial n}}{\left[\dfrac{\partial \mu}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(52)$$

For $\dfrac{\partial L}{\partial n}>0$ and for $\left[\dfrac{\partial \mu}{\partial n}\right]^{\mathrm{S}}$ and $\left[\dfrac{\partial \mu}{\partial n}\right]^{\mathrm{F}}$ of the same sign, increasing the temperature reduces the amount of the common component absorbed by the solid, (i.e. reduces the amount of swelling), but increases the amount absorbed by the liquid, i.e. increases solubility.

In a similar manner, we have for variation of the composition with stress

$$\frac{\partial n}{\partial X_{kl}}=\frac{-\dfrac{\partial p}{\partial X_{kl}}}{\dfrac{\partial p}{\partial n}}, \qquad \ldots\ldots(53)$$

and for variation of temperature with stress

$$\frac{\partial T}{\partial X_{kl}}=\frac{-\dfrac{\partial p}{\partial X_{kl}}}{\dfrac{\partial p}{\partial T}}. \qquad \ldots\ldots(54)$$

If there is only one component in each phase, both $\dfrac{\partial V^{\mathrm{S}}}{\partial n}$ and $\dfrac{\partial V^{\mathrm{F}}}{\partial n}$ are specific volumes. With this interpretation, equation (54) can be used to obtain the stress coefficient of melting temperature of a pure substance.

The difference in derivatives for the case of solution when this takes place off a stressed face or a stress-free face is interesting. In the latter case, the effect of stress on the free-energy change is only manifest in the loss of the strain energy of the material transferred to the fluid phase. Strain energy is positive for positive or negative stresses, and so the effect on the independent variable, e.g. the pressure on the fluid phase, is independent of the sign of the stress. In the case of solution from a stressed face, the potential energy of the external forces is changed and the change is of opposite sign for tension and compression; hence the effect on the pressure due to this effect depends on the sign of the external forces.

It shoud be noted that all the terms in the expressions obtained are functions of the independent variables. For example, the term $\dfrac{\partial e_{11}}{\partial n}$ in expression (26) is a function of temperature, of concentration and of the stress. It is not sufficiently accurate to substitute for it its value at zero stress. Some idea of its stress dependence may be obtained as follows. If we divide the strain into a strain due to swelling in the absence of stress ($=e_{sw}$) and a strain due to stress in the absence of swelling ($=e_{st}$) and if we neglect interaction of swelling and stress, we may write

$$\begin{aligned} e &= e_{sw} + e_{st} \\ &= e_{sw} + \frac{X}{E}, \\ \frac{\partial e}{\partial n} &= \frac{\partial e_{sw}}{\partial n} - \frac{X}{E^2}\frac{\partial E}{\partial n}. \end{aligned}$$

For spruce under tension along the grain and at about 15% moisture content, taking the unit for n as 1% of weight of dry solid, $\frac{\partial e_{sw}}{\partial n} \approx 0{\cdot}0001$, whereas at 20,000 lb./sq. in. stress $-\frac{X}{E^2}\frac{\partial E}{\partial n} \approx 0{\cdot}00035$, a value three and a half times $\frac{\partial e_{sw}}{\partial n}$. For a fabric-reinforced plastic, stressed to 15,000 lb./in.² in the plane of the laminations, similar figures are $\frac{\partial e_{sw}}{\partial n} = 0{\cdot}001$ and $-\frac{X}{E^2}\frac{\partial E}{\partial n} = 0{\cdot}0005$. In compression e_{st} has opposite sign to e_{sw}, and it is quite possible that $\frac{\partial e}{\partial n}$ may be negative at high compression stresses. In a similar manner $\frac{\partial e_{11}}{\partial T}$ in expression (38) will be affected by the change in Young's modulus with temperature.

The derivatives given in §§ 3.1–3.3 are obtained by exact thermodynamic analysis and include terms omitted by other authors. For example, Barkas (1945), by using the method of thermal cycles, has derived an expression for $\frac{\partial p}{\partial X_{11}}$ similar to that in equation (26) but his expression omits the first two terms of the numerator. It is therefore of interest to estimate the error due to this. The following figures are for a typical fabric-reinforced bakelite material swelling due to water absorption from a vapour phase. The swelling coefficients are estimated for a stress of 15,000 lb./sq. in. $\frac{\partial e_{11}}{\partial n} = 0{\cdot}0015$, $\frac{\partial e_{22}}{\partial n} = 0{\cdot}001$, $\frac{\partial e_{33}}{\partial n} = 0{\cdot}01$, the unit for n being 1% change in moisture content estimated on dry weight. $\frac{\partial e_{11}}{\partial X_{11}} = 6{\cdot}7 \times 10^{-7}$, $\frac{\partial e_{22}}{\partial X_{11}} = -3{\cdot}35 \times 10^{-7}$, $\frac{\partial e_{33}}{\partial X_{11}} = -1{\cdot}67 \times 10^{-7}$, the unit for X_{11} being 1 lb./sq. in. At a stress of 15,000 lb./sq. in. the sum of the first two terms of the numerator of equation (26) is about 8% of the third term, so that in this somewhat extreme case Barkas's expression is in error by some 8%. For spruce subject to tension along the grain the following figures are estimates of values at 20,000 lb./sq. in. stress, the units for n and X being as before. The moisture content is about 15%.

$$\frac{\partial e_{11}}{\partial n} \approx 0{\cdot}00045, \qquad \frac{\partial e_{22}}{\partial n} \approx 0{\cdot}003, \qquad \frac{\partial e_{33}}{\partial n} \approx 0{\cdot}0015.$$

$$\frac{\partial e_{11}}{\partial X_{11}} = 1 \times 10^{-6}, \qquad \frac{\partial e_{22}}{\partial X_{11}} = -4{\cdot}5 \times 10^{-7}, \qquad \frac{\partial e_{33}}{\partial X_{11}} = -5{\cdot}4 \times 10^{-7}.$$

At a tensile stress of 20,000 lb./sq. in. the sum of the first two terms of the numerator of equation (26) is about 20% of that of the third term. This is of course an extreme case, as the stress chosen is about equal to the breaking stress of the wood. The error is proportional to the stress.

For spruce subject to simple shear along the grain (see equation (30)) $\frac{\partial e_{12}}{\partial X_{12}}$ has a value of about 2×10^{-5} at a stress about equal to the shear strength of the material (taken as 1000 lb./sq. in.). The value of $\frac{\partial V}{\partial n}X_{12}\frac{\partial e_{12}}{\partial X_{12}}$ in equation (30) is therefore about equal to the value of $\frac{\partial V}{\partial n}X_{11}\frac{\partial e_{11}}{\partial X_{11}}$ of equation (26). If we take $\frac{\partial e_{12}}{\partial n}=-\frac{X_{12}}{G^2}\frac{\partial G}{\partial n}$, and if we assume that $\frac{1}{G}\frac{\partial G}{\partial n}$ equals $\frac{1}{E}\frac{\partial E}{\partial n}$, then the value of $\frac{\partial e_{12}}{\partial n}$ at 1000 lb./sq. in. shear stress is about equal to the value of $\frac{\partial e_{11}}{\partial n}$ at 20,000 lb./sq. in. tensile stress. The effect of shear on vapour pressure in this case is of the same order as that of tensile stress in the direction of the grain.

If spruce is subject to a tension at right angles to the grain, the following figures apply:

$$\frac{\partial e_{33}}{\partial X_{33}}\approx10^{-5},\qquad \frac{\partial e_{11}}{\partial X_{33}}\approx-0{\cdot}03\times10^{-5},\qquad \frac{\partial e_{22}}{\partial X_{33}}=-0{\cdot}55\times10^{-5}.$$

At a tensile stress of 500 lb./sq. in. (of the order of the breaking stress) the error in omitting the first two terms of the denominator of expression (26) is of the order of 1%. For this case Barkas's expression is sufficiently accurate.

The change in equilibrium moisture content with stress may be estimated using equation (53). For spruce at 18° C. and in equilibrium with 60% relative humidity and subjected to tensile stress along the grain, the proportional change in moisture content per lb./sq. in. stress $\left(\frac{1}{n}\frac{\partial n}{\partial X_{11}}\right)$ is about $0{\cdot}1\times10^{-6}$ at zero stress, rising to about $0{\cdot}5\times10^{-6}$ at 20,000 lb./sq. in. tensile stress. To calculate the latter figure, the stress-free value of $\frac{\partial p}{\partial n}$ has been used. The value of $\frac{\partial p}{\partial n}$ at 20,000 lb./sq. in. may differ somewhat from this. For the fabric-reinforced bakelite, the value of $\frac{1}{n}\frac{\partial n}{\partial X_{11}}$ is estimated to be about $0{\cdot}2\times10^{-6}$ at zero stress, rising to about $0{\cdot}3\times10^{-6}$ at 15,000 lb./sq. in. tension in the direction of the laminations.

§4. DERIVATIVES OF ENTROPY FOR SOLID-FLUID EQUILIBRIUM

The discussion is confined to equilibrium between one- and two-component phases. The expressions in 3.2 for the change in pressure on the fluid phase with temperature involve the latent heat of transfer of the common component from the solid phase to the fluid phase. It is therefore of interest to compute the derivatives of latent heat with respect to the explicit variables of the system. There is, however, a difficulty in obtaining results having much generality. When one of the explicit variables of the system is varied, another must be sumultaneously varied to ensure equilibrium. Thus it is not possible to compute true partial derivatives of latent heat. For example if, when the stress on the solid phase is varied, the pressure on the fluid phase is varied to maintain equilibrium, we have

$$\frac{\partial\left(\dfrac{\partial L}{\partial n}\right)}{\partial X_{kl}} = \frac{\partial}{\partial X_{kl}} T\left(\frac{\partial S^{\mathrm{F}}}{\partial n} - \frac{\partial S^{\mathrm{S}}}{\partial n}\right) \qquad \ldots\ldots(55)$$

$$= T\left[\frac{\partial}{\partial p}\frac{\partial S^{\mathrm{F}}}{\partial n}\frac{\partial p}{\partial X_{kl}} - \frac{\partial}{\partial X_{kl}}\frac{\partial S^{\mathrm{S}}}{\partial n}\right]. \qquad \ldots\ldots(56)$$

On the other hand if, to maintain equilibrium, we altered the composition of the fluid phase, we would obtain

$$\frac{\partial\left(\dfrac{\partial L}{\partial n}\right)}{\partial X_{kl}} = T\left[\frac{\partial}{\partial n}\frac{\partial S^{\mathrm{F}}}{\partial n}\frac{\partial n}{\partial X_{kl}} - \frac{\partial}{\partial X_{kl}}\frac{\partial S^{\mathrm{S}}}{\partial n}\right], \qquad \ldots\ldots(57)$$

or again, we could alter the temperature or another of the stresses on the solid phase. If the latter, then we obtain (calling the second stress which is varied X_{mn})

$$\frac{\partial\left(\dfrac{\partial L}{\partial n}\right)}{\partial X_{kl}} = T\left[-\frac{\partial}{\partial X_{kl}}\left(\frac{\partial S}{\partial n}\right)^{\mathrm{S}} - \frac{\partial}{\partial X_{mn}}\left(\frac{\partial S}{\partial n}\right)^{\mathrm{S}}\right]. \qquad \ldots\ldots(58)$$

In the circumstances, rather than give a multitude of formulae applying to particular cases, it seems best to compute the derivatives of entropy of the two phases with respect to their explicit variables separately. Derivatives of any particular type of latent heat may then be obtained by formulae of the type (56) to (58).

4.1. *Variation of entropy of solid phase with stress*

Equating the differentials of the first coefficient of equation (8) with respect to stress to that of the second coefficient with respect to temperature, we obtain

$$\frac{\partial S^{\mathrm{S}}}{\partial X_{kl}} = \left[\frac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial T}\right) - \frac{\partial}{\partial T}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{kl}}\right)\right]^{\mathrm{S}}, \qquad \ldots\ldots(59)$$

$$\frac{\partial}{\partial X_{kl}}\frac{\partial S}{\partial n} = \frac{\partial}{\partial n}\frac{\partial S}{\partial X_{kl}} = \left[\frac{\partial}{\partial n}\left\{\frac{\partial}{\partial X_{kl}}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial T}\right) - \frac{\partial}{\partial T}\left(V\Sigma X_{ij}\frac{\partial e_{ij}}{\partial X_{kl}}\right)\right\}\right]^{\mathrm{S}}. \qquad \ldots\ldots(60)$$

For simple direct stress this becomes

$$\frac{\partial}{\partial X_{11}}\frac{\partial S^{\mathrm{S}}}{\partial n} = \frac{\partial}{\partial n}\left[\frac{\partial V}{\partial X_{11}}X_{11}\frac{\partial e_{11}}{\partial T} + V\frac{\partial e_{11}}{\partial T} - \frac{\partial V}{\partial T}X_{11}\frac{\partial e_{11}}{\partial X_{11}}\right]^{\mathrm{S}}.$$

For an isotropic material having shear modulus G

$$\frac{\partial}{\partial X_{11}}\frac{\partial S^{\mathrm{S}}}{\partial n} = \frac{\partial}{\partial n}\left[V\frac{\partial e}{\partial T}\left(1 - \frac{X_{11}}{G}\right)\right]^{\mathrm{S}}. \qquad \ldots\ldots(61)$$

In the case of swelling, $\dfrac{\partial}{\partial n}$ means the change in the quantity inside the square brackets due to addition of unit quantity of the common component to a large quantity of the solid.

In the case of solution $\dfrac{\partial}{\partial n}$ may be omitted if, instead of V, the specific volume $\overline{V}$ is written.

4.2. *Variation in entropy of fluid phase with pressure*

Differentiating the first coefficient in expression (22) with respect to pressure and the second coefficient with respect to temperature, we obtain

$$\frac{\partial S}{\partial p} = -\frac{\partial V}{\partial T} \qquad \ldots\ldots(62)$$

and

$$\frac{\partial}{\partial p}\frac{\partial S}{\partial n} = -\frac{\partial}{\partial n}\frac{\partial V}{\partial T} = -\frac{\partial}{\partial T}\frac{\partial V}{\partial n}. \qquad \ldots\ldots(63)$$

For the swelling case in which the fluid is a pure substance, expression (63) is simply minus the temperature coefficient of the specific volume of the fluid. For the solution case, expression (63) is minus the temperature coefficient of the selling of the fluid when unit quantity of the common component is mixed with a large quantity of fluid.

4.3 *Variation of entropy with temperature for solid and fluid phases*

$\frac{\partial}{\partial n}\frac{\partial S}{\partial T} = \frac{\partial}{\partial n}\frac{C_p}{T}$, where C_p is the heat capacity at constant pressure or stress.

For a pure phase $\frac{\partial}{\partial n}(C_p)$ means the heat capacity per unit mass, that is the specific heat. For a two-component phase it requires careful definition. To measure it, it is first necessary to find the heat required to raise the temperature of a large quantity of the phase by one degree of temperature. Then let the phase absorb unit quantity of the common component. At the initial temperature and pressure again find the heat to raise the phase by one degree. $\frac{\partial}{\partial n}(C_p)$ is the difference between these two quantities of heat.

4.4. *Variation of entropy with composition*

The variation of specific entropy with composition is $\frac{\partial^2 S}{\partial n^2}$. This can only be obtained via latent-heat measurements, and no useful thermodynamic relations expressing it in terms of other parameters which can be obtained from direct experiments are possible.

§ 5. THERMODYNAMIC RELATIONS FOR INDEPENDENT VARIABLES NOT INCLUDING STRESS

So far the independent variables of the solid phase have been temperature, composition, and stress, and for the fluid phase, temperature, composition, and pressure. Instead of stress we may use strain as independent variable. Equation (5) gives the work done in terms of strain as explicit variable. Substituting this in (1) we obtain

$$dF = -SdT + V\Sigma X_{ij}de_{ij} + \mu_1 dn_1 + \mu_2 dn_2. \qquad \ldots\ldots(64)$$

Some of the derivatives with respect to strain, similar to those with respect to stress, which are given in §§ 3 and 4, will now be given. The method is formally equivalent to that used in these sections. The independent variables are temperature, composition strain of the solid, and pressure in the fluid.

Instead of equation (24) we have

$$\frac{\partial p}{\partial e_{kl}}=\frac{\left[\frac{\partial}{\partial n}(V\Sigma X_{ij})\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(65)$$

For the cases of swelling and solution, $\partial V/\partial n$ has the significance already discussed. Instead of equation (37) we have

$$\frac{\partial p}{\partial T}=\frac{\frac{\partial}{\partial n}\left(\frac{L}{T}\right)}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(66)$$

Instead of equation (60) we have

$$\frac{\partial}{\partial n}\frac{\partial S}{\partial e_{kl}}=-\frac{\partial}{\partial n}\left[\frac{\partial}{\partial T}(V\Sigma X_{ij})\right]^{\mathrm{S}}. \qquad \ldots\ldots\ldots(67)$$

Instead of stress and strain we might also use force and displacement as independent variables. By comparing expressions (2) and (5) and (4) and (6) it can be seen that derivatives with respect to force and displacement can be obtained from those with respect to stress and strain by writing 1 for V^{S}, P_{ij} for X_{ij} and x_{ij} for e_{ij}; for example the expression equivalent to (24) is

$$\frac{\partial p}{\partial P_{kl}}=\frac{\left[\frac{\partial}{\partial n}\left(\Sigma P_{ij}\frac{\partial x_{ij}}{\partial P_{kl}}\right)-\frac{\partial}{\partial P_{kl}}\left(\Sigma P_{ij}\frac{\partial x_{ij}}{\partial n}\right)\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}} \qquad \ldots\ldots(68)$$

and that equivalent to (26) is ;

$$\frac{\partial p}{\partial P_{11}}=-\frac{\left[\frac{\partial x_{11}}{\partial n}\right]^{\mathrm{S}}}{\left[\frac{\partial V}{\partial n}\right]^{\mathrm{F}}}. \qquad \ldots\ldots(69)$$

REFERENCES

Barkas, W. W., 1945. *Forest Products Research Special Report* No. 6.
Gibbs, J. W., 1876. *Collected Papers* (Longmans & Co.).
Warburton, F. W., 1946. *Proc. Phys. Soc.*, **58**, 585.

THE KEW RADIO SONDE

By E. G. DYMOND,
University of Edinburgh

MS. received 23 *December* 1946

ABSTRACT. The British radio sonde is a system for telemetering indications of pressure, temperature and humidity from a free balloon to the ground. It is used on a large scale for routine observations of the upper air for meteorological forecasting.

It works on the principle of a varying inductance changing the note of an audio-frequency oscillator, which modulates the radio transmitter. The design of airborne instrument, ground receiving apparatus and calibrating plant is described. An account is given of the performance of the radio sonde, and of the errors to which it is subject in actual operation. The probable errors are in the neighbourhood of ±5 mb. and ±0°·4 c. for pressure and temperature over the atmospheric range up to 22 km. height, and ±10% relative humidity down to temperatures of −20° c., below which the hygrometer element becomes unreliable or inoperative. The reliability is high, over 95% of the soundings being successful.

---

§ 1. INTRODUCTION

A RADIO SONDE is a meteorological instrument which can be attached to a free balloon in order to measure pressure, temperature, and humidity during ascent. The indications are transmitted to the ground by a radio link. A network of stations using such instruments enables a three-dimensional picture of the atmosphere to be obtained, thus providing the forecaster with far more information than can be derived from surface measurements alone.

The following is an account of the radio sonde developed for the Meteorological Office during the war. It is generally known as the Kew Mark IA radio sonde. It has been briefly described, in a general review of recent meteorological developments, by Johnson (1946).

The requirements for such an instrument are that it shall measure pressure and temperature to an accuracy of at least 1% of the range, humidity to between 5 and 10%, that it shall be sufficiently light to be carried by a balloon to a height of 16 km. on a majority of occasions, and that its cost shall be sufficiently low to allow of large-scale use, even when the chance of recovery after a flight is small.

The progenitor of the Kew radio sonde was an instrument designed by Thomas (1938) at the National Physical Laboratory. This gave continuous readings of pressure and temperature but there was no means of measuring humidity. It incorporated two audio-frequency oscillators with variable inductors, each of which was controlled by a meteorological element in such a way that the frequency of oscillation was a function of pressure or of temperature. The two oscillators simultaneously modulated a radio transmitter. Reception on the ground was by a normal communications receiver, the output from which was matched in frequency by ear with that of a calibrated variable oscillator. The two audio-frequencies were sufficiently spaced so that no confusion arose between them. Power for the balloon-borne instrument was provided by dry cells.

This radio sonde was not satisfactory for two reasons. No measurement of humidity was possible, nor was the frequency stability adequate to give acceptable accuracy. A modified form was produced by Thomas in 1939 in which the inductors were redesigned and a humidity unit added. The most important change was the substitution of mumetal for silicon iron as the material of the inductor cores. The frequency stability was much increased and adequate accuracy in the air could be expected. But the instrument still suffered from the following disadvantages:—

1. It was very heavy, weighing 2920 grams, of which the battery represented 960 grams.
2. The temperature unit, though possessing the desirable feature of very quick response to changes in temperature, was insufficiently stable. In particular it was unduly sensitive to gravity and to changes in tilt. As a radio sonde

is apt to swing through quite large arcs while ascending, unpleasantly large errors would be introduced.

3. The humidity unit was unsuited to routine use. It operated by causing a change in the pressure calibration in steps as the humidity varied.

For the above reasons the Thomas instrument was not adopted for use in the Meteorological Office, but a new design was called for, operating on the same general principle of variable inductance, but avoiding the undesirable features described above.

General features of design

As the humidity unit of Thomas was quite impracticable when applied to an instrument required for large-scale use, owing to the large number of separate pressure calibrations required, it was decided to add a third inductor, similar to those for pressure and temperature. But if a third audio-frequency circuit were added, the weight and complexity of the instrument would be increased to an inadmissible extent. Furthermore, reception would become difficult, as three audio-frequencies would be transmitted instead of two. Trials with an experimental model incorporating three oscillators confirmed this view and showed that this line of attack was not feasible.

Accordingly, in the Kew instrument, there is only one audio oscillator. Each inductor is in turn connected into the circuit by means of a switch which is driven by a wind vane. The vane rotates in the vertical slip stream. Two good features of the Thomas design were sacrificed, continuous recording and total lack of moving parts other than those of the meteorological elements themselves; but it now became possible to measure humidity in a simple manner and also to reduce the number of valves in the circuit. As an indication of the extent of simplification, the total weight was reduced from the 2920 grams of the Thomas model to 1400 grams. This, however, was partly achieved by the adoption of another type of battery.

The individual meteorological elements, with their inductors, form a unit which is detachable, as in the Thomas instrument. This is of importance in calibration, as the test chambers required are very much smaller than if the whole radio sonde had to be placed within them. In order to cheapen production the three units were made as nearly as possible alike.

§ 2. DESCRIPTION

A schematic diagram of the radio sonde is given in figure 1, which shows the three variable inductors operated respectively by an aneroid capsule for pressure, a bimetallic strip for temperature and a strip of goldbeater's skin for humidity. A photograph of the complete instrument without its container or battery is shown in figure 2. The battery rests on the lower circular panel. A cylindrical case of bakelized cardboard protects the two panels. The three inductors, which project from the case into the air stream, are shown in their shields to protect them from solar radiation. In figure 3 is given the circuit diagram. All three valves are of type HL 23, taking 55 ma. filament current.

The audio oscillator

This consists of a Hartley oscillator (V_1) with a frequency range of 700 to 1000 cycles per sec. C_1 is the condenser, of 0·07 mF. capacity, which with the inductor in circuit at the moment forms the tank circuit of the oscillator. The precision of the instrument depends largely on the stability of this condenser. As an overall constancy of 0·5 c./s. in frequency is aimed at, this condenser must maintain its capacity to well within one part per thousand even at the lowest temperatures likely to be reached. Silvered mica condensers were first used, but difficulties in production of the necessary quantities led later to the adoption of a clamped mica type* in which a very low temperature coefficient of capacity could be achieved. The average value of this coefficient in production samples is 29×10^{-6} per °c., though figures of 10×10^{-6} are possible in selected condensers. Many types of silvered mica condenser have temperature coefficients up to 50×10^{-6} per °c.

It is obviously important that the frequency should change as little as possible with battery voltage. A large measure of stabilization has been achieved by the method adopted by Thomas, in which a combination C_2R_2 is inserted between the oscillating circuit and the driving valve. The large impedance of this combination tends to swamp small changes in valve constants which might cause frequency variation. Frequency instability due to changes in grid current is discussed in a later section. Values of C_2 and R_2 are chosen empirically to give the best performance. It is possible to reduce the variations due to changes in anode voltage to zero over a small range, but those due to low-tension changes cannot be compensated at the same time. A compromise is arrived at in which simultaneous alteration of high and low tension supplies causes frequency variations of opposite sign. Typical figures are $-0{\cdot}12$ c./s. per volt for high tension and $+3{\cdot}5$ c./s. per volt for low tension changes. In actual practice this means an overall variation of $+0{\cdot}3$ c./s., due to the average drop in battery voltage to be expected during a flight.

An important component of the circuit is the condenser C_9. This serves to decouple the grid of V_1 from radio-frequency oscillations arising in the transmitter. Without this condenser, it is found that there is some feed-back through the modulator stage. The radio-frequency voltage appearing on the grid of V_1 is rectified and biases it back sufficiently to alter the audio frequency. This effect is particularly disturbing as it depends in magnitude on which inductor is in

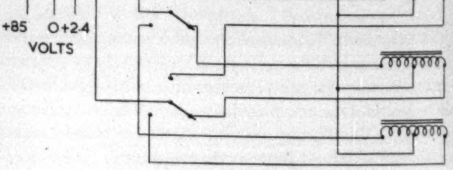


Figure 3. Circuit diagram.

$R_1=R_2=R_5=R_6=22$ K.
$R_3=R_4=47$ K.
$C_1=0{\cdot}07$ mF.
$C_2=0{\cdot}01$ mF.
$C_3=C_4=0{\cdot}001$ mF.

$C_5=5-20$ pF.
$C_6=20$ pF.
$C_7=5$ pF.
$C_8=100$ pF.
$C_9=500$ pF.

---

* Designed by the Dubilier Condenser Co. Ltd.

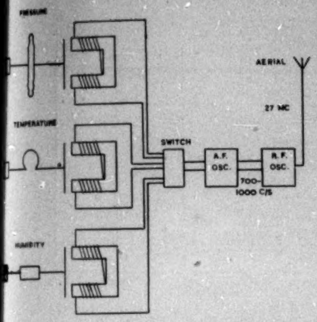

PRESSURE
TEMPERATURE
HUMIDITY
SWITCH
A.F. OSC.
R.F. OSC.
AERIAL
27 MC
700-1000 C/S

Figure 1. Schematic diagram of the Kew radio sonde.

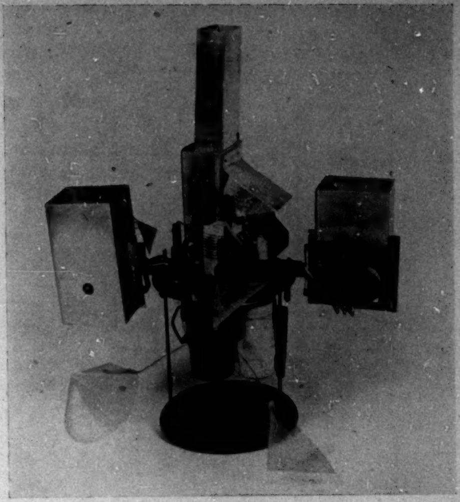
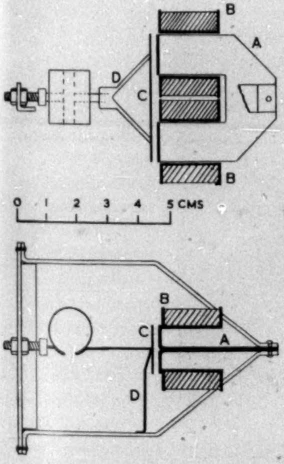


Figure 2. The radio sonde, without container or battery.


A
B
B
C
D
0
1
2
3
4
5 CMS
B
C
A
D

igure 4. An inductor, with temperature element. Plan and elevation. Slightly diagrammatic.

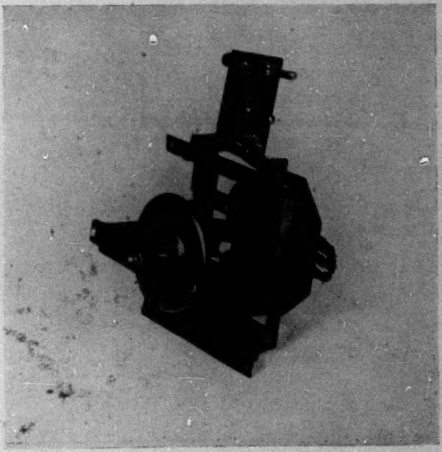


Figure 5. The pressure unit, without radiation shielding.

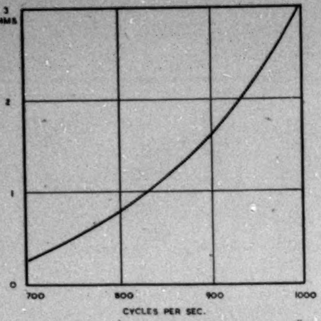


Figure 6. Relation between width of gap in the magnetic circuit and oscillator frequency.

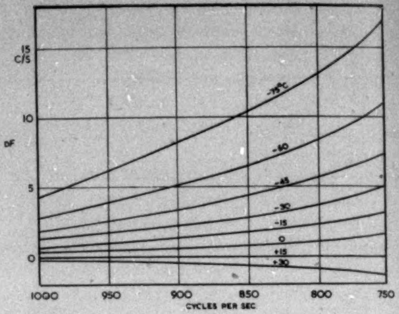


Figure 7. Frequency change, *dF*, against oscillator frequency, for various temperatures.

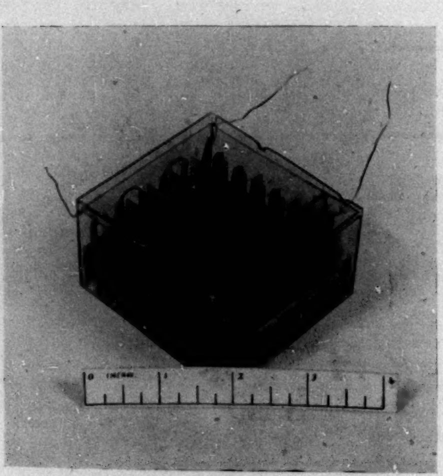


Figure 8. The battery, providing 86 volts high tension and 2·4 volts low tension.

circuit, due to differences in the r.f. impedance of the leads, and on the loading of the r.f. stage. A decoupling condenser of 500 pF. capacity is sufficient to make all variations due to r.f. interaction less than 0·1 c./s.

During calibration, only the inductor units are placed in the cold chamber. It is therefore necessary that the remainder of the oscillator circuit as a whole should be insensitive to temperature, as during an actual ascent it will be subjected to cold conditions. This question has been studied by reversing the normal calibrating conditions. The body of the instrument is placed within the calibrating chamber, with the inductor outside at room temperature, connections being made with leads through the chamber wall. It is found that a temperature change from +15 to −60° c. causes a frequency shift of +0·4 c./s. As the condenser C_1 alone should give about +1·0 c./s. in these conditions, the rest of the circuit, including the valve, produces a shift of about −0·6 c./s. The net effect is sufficiently small to be neglected.

Modulator stage

The valve V_2 is interposed between audio- and radio-frequency oscillators to impose a constant load on the former and so maintain its stability. Some measure of amplification also occurs in this stage.

Radio transmitter

This works in the frequency range 26 to 30 Mc./s. The single valve oscillator V_3 is grid-modulated by the output of V_2. The depth of modulation can be controlled by the magnitudes of the coupling resistor and condenser, R_5 and C_4, between V_2 and V_3. It is found that two states of modulation are possible. In one, the modulation depth is 30% or less, with little distortion of the audio-frequency wave-form. In the other, the oscillator V_3 is over-modulated, as during part of the cycle it is completely cut off. Depths between 30 and 100% cannot be obtained. The state of over-modulation was chosen to obtain the maximum signal strength. The wave-form is much distorted, but this is of no consequence.

With this type of circuit there is a large measure of frequency modulation in addition to the amplitude modulation. This has not been accurately measured, but the frequency swing is of the order of +50 Kc./s. When receivers with narrow pass bands are used, there is little gain in signal strength when the modulation depth is increased from 30 to 100%. But in addition to its use for measuring pressure, temperature and humidity, the radio sonde is also used for measuring winds, as described by Johnson. The direction-finding receivers used for this work are not as selective as those for radio sounding and the mode of greater modulation gives some advantage in signal strength.

The aerial is end-fed, half a wave-length long, and is attached alongside the suspending cords from the balloon. The aerial current as measured by a thermal milliammeter inserted in the midpoint of the aerial is of the order of 15 to 25 ma. It depends largely on the particular valve, as these vary widely in the efficiency of oscillation. In practice, the radio-frequency valve is selected to give a minimum output of 15 ma., with an anode voltage of 85 volts. Assuming an aerial impedance of 70 ohms, these currents correspond to a radiated power of 16 to 44 milliwatts. This power is amply sufficient. The maximum altitude of ascent is reached in

45 to 60 minutes, and it is only on very rare occasions that the instrument is carried by the wind more than 100 miles (160 km.) in this time. But instances have been recorded of the radio sonde from one station being heard by another when over 175 miles (280 km.) distant. It appears that limitation in range is due rather to passing of the transmitter below the horizon than to fading of the signal owing to distance alone.

As the radio transmitter is subjected to a large change in temperature during flight, its frequency suffers a steady drift. This is minimized by use of an air-dielectric condenser C_5 as the tuning condenser. Drifts are also caused by the voltage drop of the battery. In practice the frequency shift during a flight averages 100 and rarely exceeds 200 Kc./s.

Meteorological units

These, as already mentioned, are separate and can be plugged into the main body of the instrument. Whereas the battery and circuit elements are enclosed for protection against the weather and for thermal insulation, the meteorological elements and their inductors are fully exposed to the air. This is of importance as the inductors are somewhat sensitive to temperature, and in order to apply corrections accurately their temperature must be known with some precision. The inductors are all alike, with the exception of that in the humidity unit, in which the number of turns of the coils is reduced. This is for the purpose of slightly raising its frequency range, so as to separate it to some extent from those of the other two units.

It is important in the design of these units that they shall be mechanically robust, so that they do not distort under the shocks of transport, and that there should be a minimum of metal in regions where stray magnetic flux is strong. This flux sets up in all the metal parts eddy currents, which react very unfavourably on the frequency stability.

An inductor is shown diagrammatically in plan and elevation in figure 4. A photograph of the pressure unit without radiation shield is given in figure 5. The core A is made of six mumetal stampings, each 0·005 in. (0·127 mm.) thick, whose ends are turned up to form flat pole pieces. The coils B are wound on moulded formers, each with 1200 turns of No. 38 S.W.G. copper wire, insulated with fused cellulose acetate. In the case of the humidity unit, the coils are wound with 1100 turns only. The coils are placed as close to the pole pieces as possible, to reduce the leakage flux. The resistance of the two coils is about 120 ohms at 15° c., but falls to 80 ohms at $-60°$ c.

The moving armature C is a single stamping of mumetal 0·005 in. (0·127 mm.) thick, supported by a nickel silver stamping D of the same thickness. One portion of this stamping serves as a spring hinge for the armature; to the other is attached the meteorological element. The flux density in the armature is much greater than in the core, and it is necessary that it should be of the highest permeability. This is required not so much to obtain a high value of inductance as to reduce the losses in the mumetal. These losses, partly due to hysteresis and partly to eddy currents, are smaller the higher the permeability. It is therefore important that the strip from which the armature is cut should have been rolled in the direction of its long axis, in order to obtain the best magnetic qualities, and that after heat treatment the armature should be treated carefully. Mumetal in its high permeability state is very sensitive to mechanical handling. It is found that the stamping D may be soldered on without affecting the permeability, but that spot welding is definitely deleterious.

It will be noted from figure 3 that grid current flows through one coil of the inductor. This current, though small, produces a permanent flux in the magnetic circuit which reduces the incremental permeability. Any change in its value will therefore alter the oscillation frequency. An attempt was made to improve the frequency stability by using a grid leak with condenser coupling, thus removing the grid current from the coil. It was found, however, that the stability was worsened. Apparently when the supply voltages are altered the effect of changing grid current partly compensates for the variations in the other characteristics of the valve.

A typical curve showing the variation of frequency with air gap in the magnetic circuit is given in figure 6. The useful range of movement of the armature is about 2·5 mm. The sensitivity varies from 40 c./s. per mm. at 1000 c./s. (wide gap) to 230 c./s. per mm. at 700 c./s. But the precision of measurement cannot be expected to increase in the same proportion, as at the low-frequency end the gap is so small that slight distortions in the frame of the unit will have proportionately a bigger effect. The design is such that meteorological conditions on the ground correspond to high frequencies, both for pressure and temperature.

The pressure unit

The sensitive element is an aneroid capsule of steel, K monel or beryllium copper. A very low value of elastic hysteresis is required, and also a nearly linear relation between deflection and pressure. The Meteorological Office specification calls for a maximum width of the hysteresis loop of 2 millibars in a complete pressure cycle of 1000 mb. amplitude.

A deflection linear with pressure, when combined with a frequency-deflection relation as shown in figure 6, gives a very desirable characteristic to the instrument. Due to the logarithmic relation between atmospheric pressure and height, a much higher pressure sensitivity is required at low than at high pressures. The Kew radio sonde does not achieve the ideal of linear height sensitivity, but approaches it more nearly than do most such instruments.

The pressure unit is sensitive to temperature and, as it is subjected to the full range of atmospheric temperature, it is necessary to evaluate corrections for this. The change dF in frequency due to temperature may be expressed as

$$dF = dF_1 + dF_2 + dF_3,$$

where dF_1 is due to the aneroid capsule itself, dF_2 to the inductor coils, and dF_3 to the mumetal in armature and core. dF_1 is caused by the temperature coefficient of the elastic constant of the capsule. The effect of this is greatest at ground level, when the stress on the capsule is a maximum, and it becomes zero at the top of the atmosphere. The temperature changes at ground level are, however, comparatively small, and so the importance of dF_1 increases with height to a maximum at about 200 mb. pressure and decreases again with further increase in height. It is possible to introduce some form of compensation in the capsule, but this course has not been followed as it is found that the contribution of dF_1 to the whole effect is small.

The resistance of the coils changes by about 30% between +15 and −60° c. This alteration in the resistance of the oscillating circuit changes its frequency by dF_2. dF_3 is caused by the variation in hysteresis and eddy currents in the mumetal and, in a minor degree, to the change in permeability.

The total variation, dF, is not a linear function of either temperature or frequency, and it also varies widely from instrument to instrument. The average values are shown in figure 7, where dF is plotted against observed frequency for various temperatures. It is seen that in the upper atmosphere, corresponding to the region below 800 c./s., the corrections which must be applied are quite large and may reach, when converted into pressure, 20 mb. This is an important fraction of the total pressure.

Some insight into the relative magnitudes of dF_1, dF_2, and dF_3 may be gained from the following considerations :—

(*a*) By fitting an extension to hold the aneroid capsule at some distance from the rest of the inductor, the change in frequency can be observed when (1) the whole unit is cooled and (2) when only the capsule is cooled by immersion in a suitable bath. The differences give the effect of the capsule alone. Changes due to thermal contraction of the frame can be shown to be small. In a typical example it is found that $dF_1=2{\cdot}0$ c./s. at surface pressure for 80° c. change. At 200 mb. pressure (800 c./s.) $dF_1=-0{\cdot}04\ dF$, and is, therefore, negligible.

(*b*) The change in resistance of the coils on cooling is measured. Resistance is now added to the oscillating circuit when cold to restore the original value. The alteration in frequency due to the restoration of resistance is measured, and this must equal dF_2. For an 80° c. change dF_2 is found to be $-4{\cdot}8$ c./s.

(*c*) From the foregoing results dF_3 is seen to be $1{\cdot}52\ dF$ at 200 mb. It is, therefore, by far the most important factor, and its variation from instrument to instrument contributes to the wide range of dF found in practice.

The effect of the mumetal is itself complex, as both core and armature have to be taken into account. The relative effect of these components can be estimated by removing the armature altogether. But this can only give rough results, as, without the armature, the flux in the core differs widely from the working conditions.

More reliable estimates can be made from the following experiment. The whole unit is cooled to a low temperature and is then suddenly placed in an air stream of about 5 metres/sec. at room temperature, while the change in frequency with time is measured, as the unit warms up. The curve connecting dF with time consists of two exponentials, with half-value periods of 5 and 25 seconds and relative amplitudes of 3 to 1 respectively. The curve of quick decay represents the change due to the armature, which is very thin and exposed to the full ventilation. The slower curve corresponds to the coils and core, which are to be expected by reason of their construction to change their temperature together.

This experiment is of importance also in assessing the accuracy of pressure measurement. The proper application of the correction dF to observed values of frequency assumes that all parts of the pressure unit are at the same temperature or, more exactly, that they have the same relative temperature distribution during flight as during calibration. It is seen from the above discussion that the greatest contribution to dF arises from the armature, which follows the air temperature very rapidly. The coils and core have a considerable lag but do not have a large effect on dF. This lag does, however, limit the ultimate accuracy with which pressure can be measured, which, as will be shown later, is of the order of 5 mb.

The temperature unit

The sensitive portion is a bimetallic strip, 0·025 mm. thick, rolled into a cylinder 1 cm. diameter by 1·6 cm. high. The bending into a circular arc sets up stresses which are only slowly relieved, in spite of repeated cycling of the element over the complete working range of +30 to −70 °c. Investigation has shown that the rate of recovery with time after bending is exponential. In the first instruments, a bimetal of brass and invar steel was used, in which the amount of creep after rolling equalled the deflection due to 3°·25 c. change; 90% of the deflection was reached in 135 days. It is not possible to anneal this material effectively, as high temperatures lead to softening of the brass and destruction of the elastic properties. A combination of ordinary and invar steel was finally adopted, which gives a creep after rolling equivalent to 1°·3 c., with a 90% period of 110 days. Although this type of bimetal can be annealed, this is not possible in the actual elements used. These are arranged to curl up with increase of temperature, and at the annealing temperature the edges of the split cylinder would meet and set up fresh stresses. In spite of having only three-quarters of the sensitivity of the brass-invar combination, the new material has been found to give markedly improved performance, due to its better elastic properties. As calibration takes place at least one and usually several months after rolling, the effect of creep can be kept quite small.

At low temperatures, the sensitivity of bimetal is reduced, owing to the increase in Young's modulus. This is countered by the increasing sensitivity of the inductor for small gaps, as shown in figure 6. The combination gives a variation of frequency with temperature which may be made linear or slightly increasing with falling temperature, depending on the sensitivity of the particular bimetal.

An important characteristic of a temperature element for radio sonde work is its speed of response. In this instrument the time for a 50% response to an instantaneous temperature change is 4·5 sec. in an air stream of 5 m./s. and of normal density. The thickness of metal forming the bimetallic strip is the controlling factor in determining the lag. It is not feasible to reduce this further without impairing stability, as the cylinder would become too weak. It is also important that the strip should be ventilated as freely as possible. The lag gives rise to a systematic error near the ground of 0°·2 c. when the instrument is rising at its normal rate in an atmosphere with a lapse rate of 6° c. per km. This error varies inversely as the square root of the air density, and will amount to 0°·45 c. at 200 mb., about the level of the tropopause. In the stratosphere the error becomes negligible, since in this region the temperature itself is practically constant.

No attempt is made to apply a correction for the effect of lag. Its variation with height leads to a very small but significant error in the lapse rate; however, since all instruments of the same type are equally affected, horizontal temperature gradients are not appreciably changed.

The temperature coefficient of the inductor itself is the same as that in the pressure unit. Errors due to the inductor, however, will only arise if the temperature distribution of the various parts is different in actual flight from that occuring during calibration. The performance of the instrument does not suggest that any perceptible error arises from this cause. As previously shown, the armature, which makes the largest contribution to dF, has a lag coefficient which is so close to that of the bimetal that the two may always be considered to be at the same temperature.

The most difficult problem in designing a temperature-measuring system for radio sondes is the prevention of radiation errors, particularly in the higher levels of the atmosphere, where solar radiation is most powerful and ventilation least effective. Not only direct radiation from the sun but also that from clouds beneath the instrument must be considered. At night there is also the possibility that the bimetal may be cooled by radiation into space.

On all these counts it is of the first importance that the surface of the bimetal shall be as highly reflecting as possible. Unfortunately, it has been found that a coating of nickel or chromium adequate to take a high polish is so thick that not only is the sensitivity reduced but the stability of the element is also impaired. But one commercial grade of bimetal[*] is formed of stainless steel which will itself take a high and permanent polish.

The perfect radiation shield for radio sondes has yet to be designed. There is no difficulty in protecting the sensitive element from radiation arriving at a low angle to the horizon, but the problem of dealing with high solar elevation has not been completely solved. The following requirements, some of which are mutually antagonistic, should be met:—

1. The shield must not allow solar radiation to strike the element directly.
2. It must prevent radiation reaching the element after multiple reflection within the shield.
3. The shield must not itself absorb sufficient radiation to warm appreciably the air flowing through it.
4. There must be no interference with the free flow of air past the element.

Condition 1 suggests a tall shield with narrow opening at the top, but 3 requires the reverse; 2 demands a complicated structure which will conflict with 4. The screening of the Kew radio sonde is necessarily a compromise. The present form was reached after many modifications during the last three years, and is by no means ideal, as the necessity of maintaining production continuously did not allow of drastic alterations to preceding designs. It consists essentially of a double aluminium shield which is extended in the upward direction by a thin rectangular tube (see figure 2).

It is believed that this system is fully effective in temperate regions up to the highest levels. The evidence for this statement will be discussed below, under the heading of "Performance". But in tropical regions, in the period around noon, it cannot be expected that this or any other radio sonde will give temperature readings which are not falsified by radiation effects.

The humidity unit

The sensitive element is a strip of goldbeater's skin. This material has several advantages for hygrometric measurements over the conventional hair. It is much more sensitive, it gives more reproducible readings, and its lag in conditions of changing humidity is much less. The sensitivity varies from 4 to 8% change in length, depending on the particular sample, for 100% change in relative humidity, and is independent of temperature.

As the distribution in the atmosphere of humidity with height is much more irregular than that of temperature, it is important to have an instrument with a minimum of lag. Unfortunately, the speed of response falls rapidly with falling temperature, as table 1, due to Glückauf (1947), shows.

Table 1

Temperature (° c.)	+18	0	−30	−69
Time constant (sec.)	2·4	6	60	1800

---

[*] Hiflex, supplied by Henry Wiggin and Co.

These results were obtained in an air stream of 5 m./s., about the speed of ascent of a radio sonde. The corresponding figure for hair at +18° c. is about 30 sec. Glückauf has also shown that the maximum speed of response at a given temperature occurs for changes in the neighbourhood of 50% relative humidity, and falls off both in very dry and in very damp air. In particular, the lag approaches infinity at 100% R.H. The lag at low temperatures limits the region of reasonable accuracy to above −20° c., and at −40° c. the material becomes useless for hygrometry.

Glückauf has also shown that goldbeater's skin exhibits a hysteresis effect when subjected to a cycle of humidity changes, which includes very dry conditions, but that it recovers its original calibration when it returns to above 70% R.H. There is no hysteresis between 70 and 100% R.H.

Thus it is seen that an instrument using goldbeater's skin leaves much to be desired. No other material, however, is available with better qualities, nor do the electric surface-resistance types first developed by Dunmore (1939) show any better performance at low temperatures.

The successful use of the material depends on the observance of the following practical points:

1. The skin must be single-ply and unvarnished.
2. The maximum working tension must be limited to 50 grams per cm. width.
3. After mounting on the unit, the skin must be seasoned for several hours in a saturated atmosphere, while subjected to its working tension. The material acquires a permanent strain under this treatment, without which it is impossible to obtain reproducible results.
4. In order to minimize the hysteresis, it is advisable to condition the element by placing it in a saturated atmosphere for 20 minutes both before calibration and use, and to calibrate from damp to dry conditions. This simulates the usual direction of humidity change during an ascent.
5. While after the conditioning process no further permanent change in length occurs in a saturated atmosphere, this is not true if the material is placed in liquid water. Therefore the strip of goldbeater's skin must be protected from rain. It has been found that passage through cloud does not affect the calibration, but prolonged exposure to extremely wet fog while preparing for an ascent has on occasions given rise to further stretching.

The inductor of the humidity unit is similar to those for the pressure and temperature, except that the coils are wound with 1100 instead of 1200 turns of wire. Owing to the limitation in accuracy imposed by the nature of the material, great precision of reading is not required and the range of frequency is limited to about 100 c./s., in the upper portion of the band. In the higher atmosphere, where humidity readings are useless, the frequency of the unit is then well separated from those of the pressure and temperature units.

The effect of temperature on the calibration can be neglected. Glückauf has shown that the calibration of the skin itself is independent of temperature, and corrections due to the inductor are unimportant above −20° c., where useful readings may be obtained. It is thus unnecessary to calibrate the unit except at room temperature, an important practical advantage.

The battery

The power supply for the radio sonde must have the following characteristics:—

1. Small weight.
2. Constant potential during discharge.
3. Long shelf life before use.
4. Relative insensitivity to low temperatures.

The form of battery most nearly conforming to these conditions is that used by Väisälä (1937). A similar design, of larger capacity, has been adopted for the Kew radio sonde. A photograph is shown in figure 8. Both high- and low-tension cells are constructed with lead peroxide positive and amalgamated zinc negative plates. The electrolyte is sulphuric acid, of density 1·27. After prolonged storage, the mercury on the zinc electrode diffuses into the body of the metal; to counter this effect about 1% of mercuric sulphate is added to the electrolyte. This provides a freshly amalgamated surface at the moment of use. The case is moulded in cellulose acetate.

Such a cell gives an e.m.f. of between 2·4 and 2·5 volts. The characteristics of the complete battery are shown in table 2.

Table 2

Type	Cells no.	Volts	Discharge max. (ma.)	Discharge working (ma.)	Capacity (ma. hrs.)	Cap./weight (milliwatt hrs./gm.)
HT Mk. I	36	86·0	10	6	12	7
LT. Mk. I	1	2·4	250	175	300	11
HT Mk. III	40	98·0	30	30	45	17
LT Mk. III	3	7·2	600	600	900	29

The Mark I battery is that used for the radio sonde. A single moulding contains both high- and low-tension cells. The weight complete with acid is 300 grams. This can be considerably reduced, as later developments have shown. As an example of what is now possible, figures are also given for the Mark III battery, developed for another type of instrument. The increased performance has been attained by increasing the amount of active paste relative to inactive grid in the positive plate, and by reduction of its thickness to the minimum required for mechanical strength. As the discharge rate is very high, the capacity is dependent on the surface and not on the volume of the plate.

At the working discharge rates given in the table, the e.m.f. remains constant to 5% for at least 1½ hours, which is usually sufficient not only for the ascent but for a large part of the descent also. This constancy is of assistance in maintaining the frequency stability of the oscillators.

The positive plate is given a special forming charge during manufacture, and is carefully washed and dried before assembly. The capacity falls with time, but the stated performance can still be obtained after 12 months' storage. There is evidence that most of the reduction takes place in the first three months; but if the battery is kept in a dry atmosphere, sealed from the air, there is no diminution in capacity, and in this condition the shelf life is indefinitely long.

The sensitivity to low temperature depends on the rate of discharge. The e.m.f. changes little, but the internal resistance rises, with fall of temperature. At low rates of discharge the cells can be used down to about −30° c., but at the maximum rate failure occurs at about −15° c. This has been shown by Marth (1944) to be due to precipitation of zinc sulphate from solution, leading to the formation of a high-resistance layer at the negative electrode. In actual practice the battery is well lagged in cellulose wadding and is placed within the case of the radio sonde. Its temperature during an ascent can be studied by laboratory experiments in which the thermal lag of the battery is measured in conditions closely approximating to those during a flight. These lead to the conclusion that if the maximum altitude is reached in 45 minutes, with an air temperature of −60° c., the battery will fall to −15° c. As radio-sonde failure due to battery trouble is rare, it is believed that these experiments give a pessimistic view. In many cases the instrument can be followed, after the balloon has burst, until the descent is complete, with no indication of battery failure.

The switch

This connects in turn each meteorological unit to the audio-frequency circuit. The contacts, which are of gold-plated phosphor bronze, are operated by a cam driven through worm and gear wheel by a three-armed windmill, similar to a cup anemometer. This rotates in the air stream created by the ascent of the instrument. Near the ground, the switch makes a complete cycle of operations in about 20 seconds, giving 6 sec. to record each meteorological element. In the stratosphere, the rate decreases to about 1 cycle per minute. As the amount of power available at high altitudes is very small, the switch and gear must operate with a minimum of friction, and no lubricant can be used owing to the low temperature. The contacts are protected by a closely fitting cover to prevent condensation of moisture given off by the battery.

Ground-receiving apparatus

The apparatus on the ground for receiving and analysing the signals of the radio sonde consists of

(*a*) Radio receiver.
(*b*) Calibrated variable audio-frequency oscillator.
(*c*) Cathode-ray oscillograph.

The audio-frequency signal, derived from the receiver, is applied to one of the pairs of plates of the oscillograph, and the output from the variable oscillator to the other pair. A stationary loop is seen on the screen when the two frequencies are equal. The oscillator can be set by this means rapidly to within 0·1 cycles/sec. of the frequency of the radio sonde, the value of which can be read off a dial.

The superheterodyne receiver and oscillograph are normal commercial products and call for no comment. The oscillator must be accurate to within 0·2 c./s. over the range 700 to 1000 c./s., a degree of precision which is not attained by any commercial instrument. A beat-frequency oscillator was first developed, but because of drifts in frequency it required standardizing against an electrically maintained tuning fork at short intervals. Ultimately a resistance-capacity oscillator * was used which had a reading accuracy of 0·2 c./s., and which was constant to this amount over periods of at least two hours. It was therefore only necessary to check the calibration against the tuning fork once before each ascent. The tuning fork itself is of Elinvar, but is not maintained at a constant temperature, so that variations of frequency due to extremes of temperature may be of the order of 0·1 c./s.

The procedure of taking observations and computing the results is as follows. One man observes and measures the frequency of each signal, corresponding to pressure, temperature, or humidity, as they occur in turn. He plots against time each observed frequency, a special clock graduated in 1/20 minutes being used for timing. The scales of the plotting chart are so adjusted that with readings taken to 0·5 c./s. no interpolation is required. This is necessary as the time for observing and plotting each point is only 6 sec. The record therefore consists of a series of dots, running in three lines representing frequencies of the pressure, temperature and humidity units. There is nothing to indicate which signal corresponds to which meteorological element, and as the records may intersect, it might be thought that analysis would be difficult. In actual fact this is not so, and confusion very rarely arises. At the beginning of an ascent, knowledge of the ground conditions enables the three frequencies to be identified. In general the pressure frequency is the highest. At the top of the ascent, the humidity element always has the highest frequency. At intermediate points the character of the individual records is a guide. Pressure gives a smooth curve Temperature, while running roughly parallel to pressure, will have irregularities corresponding to inversions and changes in lapse rate. Humidity may make large changes.

---

* Developed by Muirhead and Co. Ltd.

Each recording chart lasts for ten minutes. At the conclusion of this period the completed record is handed over to a computer, who applies the necessary corrections, and from the calibration charts evaluates the frequencies in terms of pressure, temperature, and humidity. The results of the ascent are fully computed and ready for telegraphic transmission some 15 minutes after the balloon has burst.

There has been no attempt at automatic registration. While the labour of the operator would be eased by a mechanical recorder, the man himself could not be eliminated. As, also, such a recorder must necessarily be somewhat complicated, the reliability of the whole installation would be reduced and maintenance, always a difficulty in remote stations, increased.

The complete cycle of pressure, temperature and humidity measurements is repeated every 20 seconds during the earlier stages of an ascent, so with the average speed of ascent of 300 m./min., points every 100 metres in the atmosphere can be evaluated.

§ 3. CALIBRATION AND CONTROL

As the meteorological units are detachable from the main body of the instrument, these can be subjected alone to varying conditions in suitable chambers. The radio sonde proper stands outside the chamber and is connected to the units within by leads. It has been established that these leads introduce a negligibly small change in frequency, provided that the audio-frequency valve is fully decoupled with respect to radio frequency.

Pressure and temperature calibrations take place in the same vessel, which holds six units, pressure or temperature, at a time. The chamber is cooled by a bath of trichlo ethylene surrounding it. The bath is vigorously stirred and can be controlled in temperature by addition of solid CO_2 or by electric heating elements. The chamber itself is ventilated by a fan, and temperatures are measured at two points at the top and bottom by means of thermocouples. The use of two couples assures the operator that there are no undesirable temperature gradients at the moment of observation. In spite of forced air circulation, there is always some difference in temperature between top and bottom of the vessel, but this should not exceed 1° c. even at the lowest temperatures. This gradient is due principally to the fact that for constructional reasons the top of the chamber is not immersed in the bath.

The usual range of temperature calibration is from +25 to −70° c. Atmospheric temperatures as low as −85° c. are occasionally encountered in this country and −90° has been recorded in equatorial regions. A number of experimental calibrations, using liquid air as the cooling agent, have established that extrapolation of the normal calibration down to these temperatures will not produce errors greater than 1° c.

The pressure unit is given a full calibration from ground pressure to 50 mb. at 15° c. It is then cooled to −65° or −70° c. and readings for 300, 200 and 100 mb. taken, giving the frequency changes dF due to the change in temperature at these various pressures. As the variation of dF with pressure and temperature is complex, and due to several causes, it is not to be expected that all units will follow the law embodied in figure 7. It is found, however, that to a first approximation the behaviour of any unit can be represented by the curves shown, when the ordinates are multiplied by a constant factor Q_f, peculiar to that unit. It is therefore only necessary ideally to determine dF at one pressure and temperature in order to find Q_f. In practice the factor is determined at three pressure points, not only to improve the accuracy but to detect the small percentage of anomalous units which do not conform to the curves of figure 7. These latter are rejected.

This procedure for applying a temperature correction to the pressure unit gives only approximate results. Higher accuracy would undoubtedly be obtained by individual exploration of each unit. Practical considerations, however, rule this out. When large numbers of instruments have to be calibrated, the extra time involved would be prohibitive. Furthermore, the computation of a sounding must be as simple and rapid as possible if the results are to reach the forecasting centre in time to be of use, The approximate corrections lend themselves to swift computation by specially designed slide rules, embodying the data shown in figure 7. This would not be possible if individual correction charts were to be used with each instrument.

Humidity calibration is carried out in a separate apparatus. This comprises a chamber holding 24 units, in which the air is rapidly circulated, by means of a blower, past ventilated dry- and wet-bulb thermometers for measuring the humidity and through one of four vessels, containing respectively warm water, saturated solutions of sodium nitrate and calcium chloride, and silica gel. Any one vessel can be selected by means of a multiple valve, so that relative humidities of 100, 70, 40 and 10% are readily obtained.

Control corrections

Before the radio sonde is used in the air, the calibration at surface values of the meteorological variables is checked in a ventilated screen. These control readings are made with the instrument ready for flight, and take place a few minutes before release.

Slight changes in frequency relative to the calibration values are usually found. These arise from a variety of causes:—

(*a*) Differing standards of frequency at calibrating and observing stations.
(*b*) Use of different battery voltages during calibration and control.

(*c*) Influence of leads and of the surrounding metal chamber of the calibration apparatus.

(*d*) Secular changes in the meteorologically sensitive elements.

Errors due to (*a*) never exceed 0·25 c./s., as each station is equipped with an electrically maintained tuning fork for the standardization of frequency. Some variation in voltage occurs from battery to battery, which in spite of the stabilization introduced into the oscillating circuit may in extreme cases give rise to 0·5 c./s. change. The effect of (*c*) is relatively constant from instrument to instrument, and lies between 0·1 and 0·2 c./s. Thus although (*a*), (*b*) and (*c*) are all small, their cumulative effect may be of importance. Under (*d*) are included changes with time of the aneroid capsule and bimetallic strip, possible distortion of the frames due to mechanical shocks, and variation in the permeability of the mumetal owing to ageing or rough treatment in transport.

The overall change in frequency due to all causes amounts on the average to 1 or 2 c./s. A constant correction, called the control correction, is applied to the calibrations. This proceeding is somewhat arbitrary as it is not to be expected that the frequency changes remain the same at all values of the gap in the magnetic circuit. One factor, (*a*), is obviously independent of gap, and (*b*) and (*c*) give rise to errors which inversely vary as the gap, while (*d*) usually increases with gap. Thus it is impossible to lay down a general rule for variation of correction with frequency, as in any particular case the relative importance of the various factors is unknown.

The application of a constant correction is therefore only an approximation, but has the merit of allowing rapid computation. In order to limit the errors that may arise from this approximation, no radio sonde is used whose control corrections for pressure and temperature exceed 5 c./s.

§ 4. PERFORMANCE

It is not easy to assess the accuracy of a radio sonde. Direct comparison with a recording meteorograph attached to the same balloon will reveal gross errors, but the performance of the meteorograph itself is not sufficiently well known for critical work. Comparison against aircraft thermometers gives some information, but as it is not possible to ensure coincidence in place and time for both instruments, a lengthy series of measurements is required, and only mean figures for accuracy are thereby obtained. It is also not possible by this means to check the radio sonde at altitudes greater than 12 km., or about 200 mb. pressure. It is also generally impossible to separate pressure and temperature errors. But pressure measurements alone may be accurately assessed by comparison of heights computed from the readings with those directly determined by radar.

Accuracy of temperature measurements

Apart from the use of aircraft comparisons, casual errors may be assessed from the following experiments. A series of ascents was made with radio sondes in which the humidity units were replaced by extra thermometer units. Each unit of a pair was separately calibrated so that the final results include calibration errors, so far as these are not systematic.

The average difference in six flights between the pair of units in each instrument is given in table 3. In view of the fluctuations, the variation of the difference with pressure is not significant. As the errors may be assumed to be equally distributed between the two elements, the probable error of one is ± 0°·4 c. In this figure are included all sources of casual error, except that due to variation of battery voltage during the ascent.

Table 3

Presure level (mb.)	800	600	470	350	250	180	100	75	50
Difference (° c.)	0·45	0·50	0·67	1·00	0·56	0·61	0·22	0·44	1·00

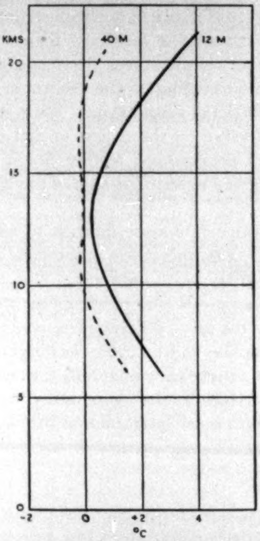


Figure 9. Difference of temperature between ascent and descent.
Full curve : Suspension between balloon and instrument 12 m.
Dotted curve : Suspension between balloon and instrument 40 m.

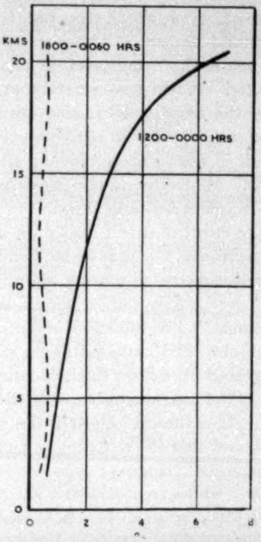


Figure 10. Mean difference of temperature between soundings at 1200 and 0000 hours (full curve), for May to July 1945, and between soundings at 1800 and 0600 hours (dotted curve).

The chief systematic errors arise from the effect of solar radiation. These are of two classes, respectively due to the action of the sun on the balloon and on the instrument itself. It is well known that solar radiation heats the balloon envelope to temperatures of which there is no reliable estimate but which may be 10 or 20° c. above that of the surrounding air. What has not been fully appreciated in the past is that the balloon as it ascends leaves a wake of heated air through which the radio sonde moves. The consequent temperature error becomes more important with increasing altitude, but can be minimized by using a sufficiently long suspension between balloon and instrument.

The error does not arise during descent, so that comparison of ascent and descent records will reveal its magnitude. Results of such a comparison are shown in figure 9. The full curve shows the mean difference between ascent and descent for a series of soundings around midday, when the distance between balloon and sonde is 12 m. The dotted curve gives the similar difference when a 40-m suspension is used. It is seen that with the shorter suspension the ascent may give temperatures too high by as much as 4°·0 c. at the highest levels, but that the error is negligible at a height of 14 km., or 150 mb. At lower levels it again becomes appreciable, but this is not due to radiation but to other factors, such as a tendency for the battery to run down towards the end of a flight, and to a pressure error, discussed below, which gives rise to a systematic difference in pressure recorded on ascent and descent. This pressure error is effective in falsifying temperatures only below the tropopause, as above this level temperatures do not vary with pressure. It will be observed that this apparent cooling on the descent at levels below the tropopause is also found with the long suspension. With a short suspension, the fall in temperature from the value on the ascent to that on the descent is nearly instantaneous and is a very pronounced feature of the flight record. When the supporting string is lengthened to 40 m. no sign of this sudden drop in temperature is found.

The effect of direct insolation of the instrument is more difficult to measure and to remove. It may be studied by comparing temperatures at the same height taken during neighbouring soundings at midday and midnight. The means taken over a long period should eliminate variations due to changing weather conditions. The full curve of figure 10 shows the mean difference in temperature at 0000 and 1200 hours during the three months May to July 1945, the observations being taken from three English stations. Such a curve shows not only the radiation error of the instrument but also any diurnal variation in the temperature of the air itself. It is indeed believed that the contribution of instrumental insolation is small, and that the diagram in fact gives a nearly true picture of the real daily temperature changes.

The evidence for this belief is as follows :—

(*a*) Many variations have been made in the form of the radiation shields around the temperature element. Also the surfaces of the bimetallic strips have been changed from a dull matt to a highly polished nickel-plated tape. This caused wide variation in the amount of radiation which the element was capable of absorbing. None of these changes has been found to make a significant change in the form of the curve in figure 10. It is reasonable to assume therefore that the instrument is insensitive to incident radiation.

(*b*) Observations are taken at 0600 and 1800 hours, and the temperature differences at these two times at various heights have been determined. A true diurnal variation will cause the air to be warmer at 1800 than at 0600 hours owing to the phase lag between air temperature and solar radiation. Direct radiation effects on the radio sonde will be equal at both times and will not appear in the difference. The results are shown by the dotted curve of figure 10. There is a difference, nearly constant with height, of about 0°·5 c. between the two times of observation, from which we can conclude that there is a true diurnal variation of temperature.

(*c*) The height of the radio sonde at any point can be computed from the pressure at that point and the temperature distribution below the instrument, using the usual barometric formula. The height can also be directly determined by radar methods to a much higher degree of accuracy. If the observed 1200–0000-hour differences (full curve of figure 10) are largely radiation errors, then in computing heights of a

> daylight sounding more accurate results would be obtained by utilizing the midnight rather than the midday temperatures. Piagsa (1946) has shown that this is not so. If the radar measurements are taken to be exact, the mean errors in height for about 50 daylight ascents are −86 metres when the temperatures as actually measured are used for computing, and +1000 metres when temperatures derived from the previous midnight soundings are taken.

We are therefore led to believe that instrumental errors due to radiation are small, and they can be roughly estimated as not exceeding 20% of the values given by the full curve of figure 10. At the very lowest pressures a larger contribution due to insolation cannot be excluded, as the number of observations at less than 60 mb. is much less than that at higher pressures.

A daily variation in the temperature of the upper air is not unexpected, though it is difficult to explain its magnitude. This question has been recently discussed by Dobson (1946).

Accuracy of pressure measurements

This can best be estimated by comparing the heights as computed from the radio-sonde observations with those directly determined by radar. It is routine practice at some stations to attach to the balloon a radar reflecting target and to observe its motion by means of a standard Army centimetric radar set, type A.A. No. 3 Mk. II. The high precision of range and elevation measurements gives an acceptable standard for checking the radio sonde.

Two series of comparisons have been made, by Harrison (1944) and by Piagsa (1945). The results are summarized in table 4, which gives not the actual height errors but the equivalent errors in pressure.

Table 4

Height interval (km.)	Radio-sonde pressure errors (mb.)	
	1944 series	1945 series
0 to 5	5·4	6·4
5 to 10	2·8	8·2
10 to 15	0·4	8·0
15 to 22	1·1	6·8

The two series made at different stations at different times are significantly different. The 1945 series is distinguished by the fact that the pressure control corrections were throughout larger than normal, for reasons not yet fully explained. These results are not therefore fully representative. The systematic errors in the lower layers in the 1944 series are partly explained by the use at that time of an incorrect temperature correction diagram. A more accurate diagram would reduce the errors up to 7 km. by about 2 mb.

Casual errors may be assessed in the same way as for temperature measurements. A series of ascents was made with instruments carrying two pressure units. The probable error of a single determination was found to rise from ±1·5 mb. at 1 km. height to a steady value of ±4·5 mb. from 13 km. upwards. The source of casual error is due to the application of the temperature correction. As previously explained, the curves in figure 7 are the means for a large number of instruments. Individual radio sondes will not follow the curves exactly. Systematic errors can arise not only when large control corrections are found, but also if the pressure unit is not at the temperature indicated by the thermometer. There is no systematic difference in pressure errors between night and day soundings, so that the radiation shielding of the pressure unit must be efficient, but as different parts of the unit make different contributions to the temperature effect and have different thermal lags, some error must arise from this cause.

This differential thermal lag is particularly apparent when heights on the ascent and descent are compared. The tropopause is found to be on the average at 10 mb. lower pressure on the descent. This is in the opposite direction from what would be expected from elastic hysteresis of the capsule, and can be attributed to the thermal lag of the mumetal core and coils of the inductance. This lag is of course only operative during the ascent, as on the descent the instrument traverses a region of practically constant temperature before the tropopause is reached. We should expect therefore that the descent should give a more accurate value of the pressure. The results shown in table 4 (1944 series), however, are not in accordance with this view, but indicate that the systematic error on the ascent is small. It is plain that there must exist another source of error, of approximately equal magnitude but opposite in sign, to that due to lag. There is not sufficient evidence to define its origin, but it seems likely that the correction curves in figure 7 are not sufficiently accurate. As they are obtained from only a small fraction of the total number of instruments used, some sampling error is to be expected.

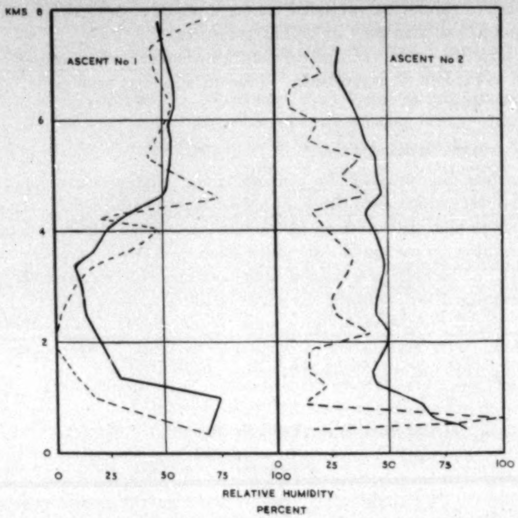


Figure 11. Comparison between humidity measurements by radio sonde (full curve) and Dobson and Brewer frost-point hygrometer on aircraft (dotted curve).

We conclude, therefore, that the pressure element is subject to casual errors of up to ± 5 mb., and while the systematic error does not in general exceed 2 mb., differences in manufacture of various batches may lead to considerably larger values if the several sources of error do not cancel.

Accuracy of humidity measurements

Here the only means of assessing systematic errors is by comparing with measurements made on aircraft ascents. Because of the low temperatures, the Dobson-Brewer frost-point hygrometer (Dobson, 1946) is the sole instrument that can be used as a standard, and it can only be used for upper-air work in an aircraft, as it requires a skilled operator.

The relative humidity of the atmosphere varies rapidly not only vertically but also horizontally. It is necessary therefore for the aircraft to keep close to the balloon during the ascent. This is very difficult to achieve in practice. As no aircraft with a sufficiently high rate of climb was available, in the trials it was necessary to operate the radio sonde at about a third of its normal ascensional velocity. The ventilation of the goldbeater's skin was accordingly deficient and the lag intensified. The result of two trials, by Harrison and Brewer (1944), are shown in figure 11. The sluggishness in response of the goldbeater's skin, especially at the higher levels, where the temperature is low, is evident. Deviations in the lowest layers between aircraft and radio sonde are to be ascribed to actual differences in the air, due to patches of clouds. The temperature differences indicated by aircraft and radio sonde on these ascents were within the probable errors and showed no anomalies in the region of gross differences in humidity.

Errors up to 25% relative humidity are shown by these trials. They do not provide a really fair indication of the performance of the radio sonde, for the reasons mentioned above, and it is to be expected that in proper conditions the instrument may attain an accuracy of better than 15% R.H.

Mutual comparison of two humidity elements on the same radio sonde, on the same lines as for pressure and temperature units, shows that the average difference is 5% R.H., with a maximum of 10%. The self-consistency of this type of element is therefore reasonably good, with a probable error of a single determination of $\pm 2\frac{1}{2}\%$ R.H.

On the other hand it can be seen that humidity measurements by radio sonde are not satisfactory. While a fair measure of accuracy is to be expected at levels in which the temperature exceeds -20° c., at lower temperatures the instrument gives little indication of the true conditions. No simple method of measuring humidity applicable to radio sondes will give acceptable results in this region. The reason for this lies in the exceedingly small quantity of water vapour which the air can contain at these low temperatures.

ACKNOWLEDGMENTS

The development of the Kew radio sonde owes much to many late colleagues in the Meteorological Office. Besides those referred to in the text, especial mention must be made of Dr. H. Carmichael, who was largely responsible for the design of the meteorological units, and of Messrs. L. G. H. Dines, I. M. Hunter, and A. J. Lander, who rendered much assistance in various stages of the work. Thanks are also due to officers in charge and to staff of the several radio sonde stations in the British Isles, who by their criticism and suggestions have done

much to improve the instrument, and whose keenness and attention is an important factor in establishing a high degree of precision.

I am indebted to the Director of the Meteorological Office for permission to publish this paper.

REFERENCES

Dobson, 1946. *Proc. Roy. Soc.*, A, **186**, 146.
Dunmore, 1939. *J. Res. Nat. Bur. Stand.*, **23**, 701.
Glückauf, 1947. *Proc. Phys. Soc.*, **59**, 344.
Harrison, 1944. *Meteorological Research Committee Reports*, MRP 203.
Harrison and Brewer, 1944. *Meteorological Research Committee Reports*, MRP 205.
Johnson, 1946. *Nature, Lond.*, **157**, 247.
Marth, 1944. Unpublished Report from the Marine Observatorium, Greifswald.
Piagsa, 1945. Unpublished Report from the Meteorological Office.
Piagsa, 1946. Unpublished Report from the Meteorological Office.
Thomas, 1938. *Proc. Roy. Soc.*, A, **167**, 227.
Väisälä, 1937. *Acta Soc. Sci. Fenn.*, **9**, 9.

THE ACCELERATION OF CHARGED PARTICLES TO VERY HIGH ENERGIES

By M. L. OLIPHANT, F.R.S., J. S. GOODEN and G. S. HIDE

MS. received 21 *March* 1947

ABSTRACT. More experimental information about the nature of the binding forces between nuclear constituents is necessary before an advance in fundamental nuclear physics can be achieved. By considering the type of information which would be most useful, the conclusion is reached that it necessary to have available protons of energies of about 1000 Mev. in order to carry out the necessary experiments. It is with a method of obtaining protons of this energy that this paper is concerned. An examination of the possibilities of achieving such high energy protons by the existing methods leads to a pessimistic conclusion, and a new method is suggested.

This new method, the synchrotron, is described in principle, and its advantages are outlined, a very important factor being its comparatively low cost. An accelerator of this type is being built at Birmingham University with a grant from the Department of Scientific and Industrial Research, and its design is considered in some detail. The magnet and its excitation form the greatest part of the apparatus in size and cost. Several alternative methods are suggested and discussed for both the magnet design and its method of excitation. An air-cored magnet is considered but rejected because of the very large mechanical forces involved and the precision required in positioning the conductors. As a result an iron-cored magnet has been chosen for construction. The excitation of the magnet is to be achieved by a d.c. motor-generator supplied with a fly-wheel. The requirements of the accelerating system, in which is included a radio frequency which changes by a ratio of about 1 : 36 during the acceleration, are quite exacting. The methods by which it is hoped that these requirements will be met are outlined. The problems associated with injection and extraction of the particles receive some attention, and a schematic description of the proposed vacuum chamber is included.

When protons of energies greater than 10^{10} ev. are to be obtained by a synchrotron, the cost of the device becomes overwhelming and some alternative method will have to be suggested. The application of the synchrotron being built at Birmingham to accelerating electrons, is limited to achieving electron energies of about 300– 00 Mev. because of radiation losses.

§ 1. INTRODUCTION

FURTHER advance in fundamental nuclear physics is dependent upon an increase in experimental information about the nature of the binding forces between the nuclear constituents. The most obvious way to obtain this knowledge is to extend Rutherford's method of exploration of nuclei by bombardment with fast particles to much higher bombarding energies than have hitherto been available and to examine the laws of scattering of protons and neutrons in very energetic collisions with similar particles.

At the present time there is no real understanding of the forces between the elementary particles and, indeed, no satisfactory explanation of the existence of only a certain limited number of such particles, with very different masses, some electrically charged, others uncharged. (Peierls, 1946). The primary problem from the point of view of the physics of nuclei is that of the proton-neutron interaction. The mass of the neutron is greater than that of the proton for reasons which are not at all understood; energetically it should be possible for a neutron to transform spontaneously into a proton and electron, but this transformation has not yet been observed. Attempts to explain proton-neutron forces in terms of virtual creation in the immediate neighbourhood of the particles of pairs of electrons or mesons or of quanta are little more than assumptions that fields of particular forms exist round them. Such attempts have failed to explain the observed binding energies of nuclei. It is unlikely that substantial progress will be made by further guessing in this field of physics unless such guesses are guided by fresh experimental facts.

Primarily, interest must centre round the interactions at close distances of approach between the elementary particles, viz. protons, neutrons, electrons, mesons. It is probable that very energetic neutrons can be produced only by bombardment of matter with high energy protons. From the practical point of view it is therefore necessary that protons and electrons should be accelerated to energies which are as high as possible and their interactions with matter observed. The relative value of protons and electrons for this purpose is difficult to determine in advance. The great success of the cascade theory of shower production in cosmic radiation suggests that the interactions of nuclei with electrons are better understood than the interactions with heavy particles.

It would appear, then, that it is essential to produce protons with energies as high as possible and somewhat less necessary to accelerate electrons to comparable energies. It is important to be quite clear about the importance of accelerating protons, for the acceleration of electrons to energies of the order of $0{\cdot}5 \times 10^9$ ev. is so much simpler and cheaper that there is a great temptation to find excuses for accelerating electrons and for postponing the difficult problem of proton acceleration.

It is necessary that observations should be carried out at energies at least equivalent to the proper energy of a pair of the hypothetical nuclear mesons, and if it is assumed that these mesons have the same mass as the free mesons observed in cosmic radiation, particles are needed with energies above about 300 Me v. Since recoiling nucleons can carry away at least half of the initial energy it is probable that bombarding particles with energies above 600 Me v. are desirable. The total binding energy of nuclei of medium atomic weight is of the order of 1000 Me v. and the character of nuclear reactions is likely to change in this region of energies.

It is clear that a good target figure at which to aim in the development of new methods of acceleration is 1000 Me v. or more. Experience of other methods of acceleration suggests that if the maximum energy for which the equipment is devised is 1000 Me v., the maximum useable energy is likely to be lower. Thus an equipment designed for 1000 Me v. will be reasonably certain to deliver energies well above 600 Me v. without straining the apparatus. To settle some questions, particles with energies much greater than 1000 Me v. may be required, but it is probable that these higher energies will be obtained only after some experience in the region of 1000 Me v. In what follows a particular method for obtaining protons with energies above 10^9 ev. is described after some consideration of reasons for preferring this method to others which have been suggested.

§ 2. LIMITATIONS OF EXISTING METHODS

Acceleration methods may be divided broadly into two classes. In the first are all systems in which the particles are accelerated along straight paths; the second includes all methods in which a magnetic field is used to bend the particles during acceleration into spiral or circular orbits.

High-voltage methods belong to the first class. Such systems possess inherent stability of particle paths, provided the ordinary rules of electron-optics are observed, but they are limited to energies less than about 10 Me v. and are troublesome above 5 Me v. To this class also belong the so-called linear accelerator methods in which the particles are pushed along by a travelling wave moving at suitable velocity along a wave-guide, or in which they pass through a series of resonators where the phases of the fields are suitably adjusted. It is an inherent defect of linear accelerators that it does not appear possible to achieve directional focusing of the particles at the same time as phase stability. It is necessary to use external focusing methods, such as an axial magnetic field, which is difficult for large apparatus and high energies, or to use thin foils across the exit openings from the accelerating gaps in the manner proposed by Alvarez.* He has commenced the construction of a linear accelerator for protons in which the exit from the gaps is covered with thin beryllium foil. He hopes to gain an energy of 1 Me v. in each foot of the accelerating system so that an apparatus to produce protons of 1000 Me v. would be about 1000 feet in length. An enterprise of this sort is practicable only when a long building can be provided or where the equipment can be used out of doors. If it is successful this equipment may provide the simplest and cheapest form of accelerator capable of extension to almost unlimited energies. There are no difficulties due to radiation from the particles as they move in straight lines and no troublesome problems of injection or extraction of the particles. However, the engineering problems are formidable and the proposed solution to the focusing difficulty is not yet proven.

One advantage of these linear methods of acceleration is that they are applicable equally to all charged particles.

Apparatus in the second class includes the cyclotron, synchro-cyclotron, betatron and synchrotron. Heavy particles, such as α-particles, have been accelerated in the cyclotron to energies of 30–40 Me v. by Lawrence and his co-workers. The relativistic increase in mass of protons at energies above about 20 Me v. makes it extremely difficult and wasteful of electric power to go to higher energies by the straightforward cyclotron method. This difficulty has been overcome by introducing a change of frequency during the acceleration (McMillan, 1945). The synchro-cyclotron is an extremely successful apparatus and promises to become the standard equipment for acceleration of protons and other heavy particles to energies of a few hundred million electron volts. However, to produce protons with an energy of 10^9 ev., a magnet is required with a field of 15000 gauss over a circular pole of radius 15 feet. Such a magnet would weigh more than 10000 tons and would be extremely expensive to build and operate. A magnet of this order of size is under construction in U.S.A., to be financed from Government funds, but it is unlikely that similar equipment can ever be available in academic laboratories.

---

* Private communication.

The induction accelerator, or betatron, of Wideroe (1928) and Kerst (1941), has been developed successfully for the acceleration of electrons, but considerations of cost and complexity render it unsuitable for the highest energies, while it cannot be used to accelerate heavy particles.

Various other attempts have been made to develop accelerating systems which can reach high energies, some of them employing resonance in a magnetic field, as those of Schwinger * and Veksler (1945), which use a combination of guiding field and linear accelerator or are modifications of the betratron, as Wasserab's "Wirbelrohr". However, none of these systems is very attractive, and they have not yet been either built or operated.

§ 3. THE SYNCHROTRON

In September 1943 one of us submitted to the Directorate of Atomic Energy in the Department of Scientific and Industrial Research, a proposal for the acceleration of electrons and protons by a new method to energies above 10^9 Me v. Subsequently, and independently, similar proposals were made by McMillan (1945) in U.S.A. and by Veksler (1945) in U.S.S.R. The name *synchrotron* was suggested by MacMillan. The essence of the new method is the conception of stable circulating orbits which increase in energy through a cyclotron type of resonant acceleration as a result of an adiabatic variation of the magnetic field, of the frequency of the accelerating electric field, or of both. The success of the synchro-cyclotron † afforded convincing proof of the validity of the general conceptions of the stability of the orbits for a system for the acceleration of heavy particles in which the frequency changes while the magnetic field remains constant. Goward and Barnes (1946) were able to demonstrate that electrons can be accelerated in a system where the radius of the orbit and the applied frequency of the electric field are constant but the magnetic field increases with time. There is a third system in which both frequency and magnetic field are varied during the acceleration. This system has been considered in detail by us and is now under construction. In what follows we give a general analysis of the proposed method and the considerations which have led to the designs adopted.

---

* Unpublished note.

† Private communications from Berkeley.

The principal practical aspect of the synchrotron method of acceleration is that for energies of the order of 10^9 ev. its cost is not prohibitive. This is due to the fact that since the orbital radius is constant, a narrow annular guiding magnetic field can be employed, so that the first cost of the magnet is much less than for a cylindrical field as used in the synchro-cyclotron. Much attention has been paid to the method of producing this field in an economical and satisfactory manner.

The essential data for the design of a synchrotron are the radius of the mean orbit, ρ, the maximum value of the magnetic field, H, and the rate of revolution, ν, of the particles in the orbit which determines the frequency of the accelerating voltage. These quantities are connected by the formulae:

$$W=\sqrt{H^2\rho^2c^2+E_0^2}-E_0, \qquad \ldots\ldots(1\,a)$$

$$\nu=\frac{c}{2\pi\rho}\sqrt{1-\left(\frac{E_0}{W+E_0}\right)^2}, \qquad \ldots\ldots(1\,b)$$

where W is the kinetic energy of the particles and E_0 is the self-energy, m_0c^2/e of the particles.

It is clear that in order to obtain high energies the maximum value of the product $H\rho$ must be large. We are concerned here with the design of a synchrotron to produce protons of energy greater than 10^9 ev., and in what follows we shall assume that W is to be $1{\cdot}3\times10^9$ ev.

The magnetic field must be so shaped that the orbits are stable (Kerst and Serber, 1941) and, in order to obtain an appreciable output in spite of inevitable radial and axial oscillations of the particles about the mean orbit, the width and depth of the annulus in which the particles move must not be too small. The magnetic field varies from almost zero to its maximum value during each cycle of acceleration, so that the construction must be such as to allow of A.C. operation. Thus the field can be generated in three ways; by using a system of conductors properly spaced and carrying appropriate currents: by using a ring-shaped laminated iron-cored magnet, with pole-pieces of the proper contour; or by shifting the magnetic flux to and from the annular space by purely electrical methods or by rotating or oscillating an electromagnet near magnetic circuits of proper design. In any case the cost of the magnet and its exciting circuits is the major item of expense and the choice of the magnet system determines all other parts of a synchrotron equipment.

§ 4. MAGNETIC FIELD AND RADIUS OF ORBIT

The variation of the energy ϵ stored in the magnetic field of a synchrotron, with the radius of the orbit, for a given ratio of gap dimensions (volume v) to radius, is given by

$$\epsilon=(H^2/8\pi)\,.\,v\sim\rho$$

for a fixed final energy of the beam produced. This magnetic field energy must be supplied by a source of electrical power, and the provision of this power represents the largest single item of expenditure. The above relation indicates that the cost of the power unit should decrease with radius and that it would pay to use the highest possible magnetic field. However, limits to H are set either by the saturation of iron or by the dimensions of, and forces upon, conductors where iron is not used.

(a) *Air-cored magnet.*—We have considered a system of conductors in the form shown in figure 1, where a current flows in one direction for conductors shown in open section, and in the reverse direction for conductors shown in solid section. If the system is straight and long compared with its diameter, a sinusoidal distribution of conductors gives a uniform field across the equator. If such a system is bent into a circular toroid the field increases across the toroid along a radius of the circle. To make the field fall off along the orbital radius, the sinusoidal distribution must be distorted, while, to enable the beam to be injected and ejected, the central conductor must be removed and suitable compensating conductors added as shown. The correct position of the conductors cannot be calculated, but a distribution can be chosen arbitrarily and the field calculated numerically. By a series of successive approximations a distribution of conductors was found which gave a reasonable approximation to the field required. However, it is found that the conductors must be placed and held very accurately in position.

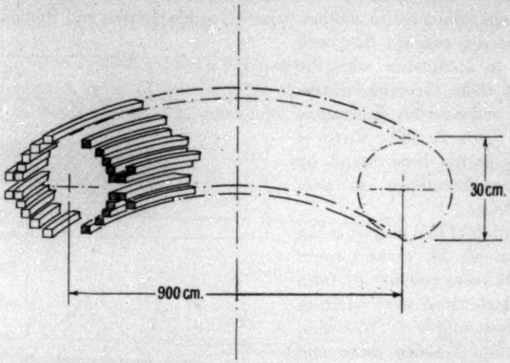


Figure 1. Arrangement of conductors for air-cored magnet.

The current in the conductors is independent of ρ for geometrical scaling, so that the relative cross-section of conductor increases as ρ decreases, giving departures from the field-form required. Also, as H increases, the forces on the conductors increase and the tolerances in position decrease, so that the problem of holding them in position rapidly becomes insuperable. Accordingly a compromise of 15 000 gauss, for which ρ is 450 cm., was chosen and the design of a synchrotron considered in detail. With a total of 22 conductors and diameter of orbital space of 30 cm., the allowable deviation of conductors from correct position is 0·1 cm., and a peak current of about 80 000 amperes is required. For operation at the equivalent of 25 cycles/sec. the peak driving potential across the coil is 25 600 volts. The forces on the conductors in this system are already of the order of the ultimate strength of copper conductors.

Continuous operation at 25 cycles would require a circulating energy of 2×10^6 kv.a. (6.4×10^6 w-sec.), while the copper losses would be 5×10^4 kw.

It is clear that such a system must be operated discontinuously by storing up energy continuously at a reasonable rate and discharging it at intervals through the coil. If the storage system is an electric condenser, a very bulky capacity battery of about 20000 microfarads is required, costing about £175000. The alternative of a short-circuit type of alternator has been considered in which the energy is stored as rotational kinetic energy, but the cost of the complete installation is of the same order of magnitude, while the engineering problems of installation and maintenance and the noise of such rapidly rotating machinery in an academic research laboratory render it even less attractive than the capacity battery.

(b) *Iron-cored magnet.*—The use of iron in a magnetic circuit reduces the volume of the useless magnetic field outside the orbital space and effects a saving of about a factor 2 in the energy stored. However, with a laminated structure the maximum flux density is limited to about 15000 gauss. The minimum dimensions of the orbital space to secure stability and a reasonable yield of particles are considered in another paper (Gooden, Jensen and Symonds, 1947). An iron-cored magnet has been designed in accordance with the results of these theoretical investigations and has the dimensions given in figure 2. The shape of the pole-tips has been found by model experiments in an electrolytic tank.

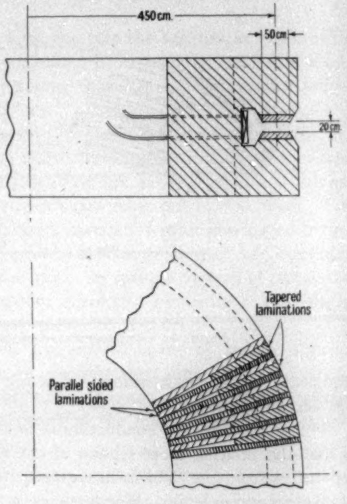


Figure 2. Proposed iron-cored magnet.

Operation of this magnet at the equivalent of 25 cycles would necessitate constructing it from laminated electrical steel which is in very short supply. Theoretical investigations (Gooden, Jensen and Symonds, 1947) indicated that the yield of protons might be improved by using a more slowly rising magnetic field and that the increased intensity of output in each pulse might compensate for a lower repetition rate. By improving the vacuum conditions it is thought that loss of particles by scattering in the residual gas can be reduced to an extent where a time of acceleration of about 1 second is practicable, with an initial injection energy of 300000 ev. A slow rate of change of field also has the advantage that the corresponding change in frequency of the accelerating potential can be more easily achieved by mechanical methods. Although the energy which must be supplied to create the magnetic field is the same as for shorter periods of excitation the power is reduced in proportion to the time of rise of field. In particular, for a time of rise of field of about 1 second, it becomes practicable to supply the energy from a d.c. generator which is provided with a flywheel in the manner used for "field-forcing" in switch-gear testing equipment.

At these very low frequencies the thickness of lamination which can be employed is large. The field is sensibly in phase over the orbital space for laminations 0·5 inch in thickness. It is thus practicable to build the magnet from rolled sheets of low-carbon steel and a very good space factor can be secured if some of these sheets are tapered. Through the cooperation of Sir A. McCance, F.R.S., of Colville's these special sheets are now in process of manufacture.

(c) *The electrical circuit.*—The relevant electrical data for the magnet are given in the following table:

Turns in winding	22
Cross-section of conductor	7 sq. cm.
Resistance	0·014 ohm.
Inductance	0·1 H.
Time-constant	7 sec.
Peak current for 15,000 gauss	11 000 amp.
Volts to give rise in 0·8 secs.	1100 volt.
Number of cycles of excitation per minute	6

The proposed cycle of operation is shown in figure 3. The generator, consisting of twin-coupled d.c. generators in parallel, driven by a 1500 h.p. motor and provided with a 36-ton flywheel, will be supplied by Messrs. Parsons, whose help we are glad to acknowledge.

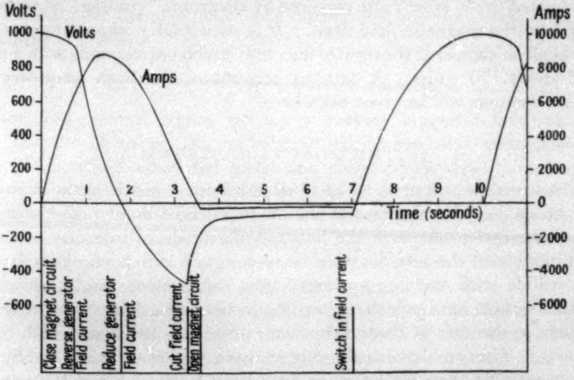


Figure 3. Cycle of operation of magnet and generator circuit.

§ 5. THE ACCELERATING SYSTEM

For the magnet under consideration the fundamental synchrotron equations (1 *a*, 1 *b*) become:

$$W = 300\sqrt{H^2(2\cdot03\times10^5)+(9\cdot61\times10^{12})} - 9\cdot3\times10^8, \qquad \text{......(2 a)}$$

$$\nu = 1\cdot06\times10^7\sqrt{1-\left(\frac{9\cdot3\times10^8}{W+9\cdot3\times10^8}\right)^2}, \qquad \text{......(2 b)}$$

where W is in electron -volts and H is in gauss.

It is clear from figure 3 that H will increase approximately linearly. The values of ν for corresponding values of H determine the frequency of the accelerating voltage applied to the accelerating electrodes. For the slow rate of acceleration chosen, the energy added per revolution of the particles is only about 200 volts and the voltage amplitude of the a.c. applied to the electrodes need be of the order of only 1000 volts. There is an optimum value for this applied voltage which leads to greatest orbital stability and maximum output current, the reason for which is discussed in another paper. The same theoretical reasoning shows that there are no advantages to be gained by using more than one electrode, the frequency applied to which is $\nu_e = \nu$. It can be shown (Gooden, Jensen and Symonds, 1947) that ν_e must equal ν to within about 0·1% over the first one-hundredth part of the acceleration, and must not differ from ν by more than about 1% thereafter.

If protons are injected at 0·3 Me v., which is about the maximum for an internal "gun", then ν_e must vary from about 0·27 Mc. to 9·5 Mc., i.e. by a factor of more than 30. A circuit of low Q, which is tuned to about 1·5 Mc., will give an adequate response when driven over the lower frequency range. A factor in frequency of about 8, which remains, can be obtained by mechanical tuning of the relatively low Q circuit through a cam of suitable shape, the frequency being adjusted to the exact value required by electronic "pulling" of the oscillator produced by the magnetic field itself. It is particularly important that the frequency shall be correct at the time of injection which corresponds with a magnetic field of about 170 gauss. A detailed account of the high frequency system of this synchrotron will be given elsewhere.

§6. THE INJECTION SYSTEM

If a reasonable output is to be obtained from a synchrotron it is essential that as many particles as possible should be injected during the "acceptance" period of the cycle. As with the betatron the mode of injection and the subsequent motions of the particles must be so designed as to ensure that the protons do not collide with the gun system during subsequent revolutions. Besides the particle oscillations which occur in the betatron, there are further radial oscillations in the case of the synchrotron, which are associated with the phase oscillations. This problem is subject to analysis (Gooden, Jensen and Symonds), though it may be that some factors have been neglected and the conclusions from the analysis may be no more applicable than similar calculations made for the betatron by Kerst and Serber (1941). However, as a result of the analysis it had been decided to place the ion source and initial accelerating system above the orbital plane and to apply a vertical electric field which will make the orbits spiral downward during injection at a rate sufficient to ensure that they miss the gun after the first revolution. Uncertainties in the analysis of the initial motions of the ions may mean that this system will need considerable modification before the maximum output is obtained. Observation of the paths of the protons after injection with steady magnetic fields, and of the paths of alpha-particles from radioactive sources with fixed magnetic fields, should make it possible to correct the injection conditions, whether arising from position of the gun or the inevitable deviations of the magnetic field from the correct form due to inhomogeneous properties of the iron at these low magnetizations.

§7. THE EXTRACTION SYSTEM

In order to ensure that the protons are obtained in a definite beam, they must be deflected by an electric or magnetic field during a single revolution. The problem is much simpler than with the betatron or with small synchrotrons, as the period of revolution in the orbit at the maximum energy is much greater, owing to the large radius of the path. A deflecting voltage can be applied to a relatively long electrode in a time short compared with the time of revolution in the orbit (10^{-7} sec.), by connecting it to a large capacity through a spark gap which is triggered to break down at the end of the acceleration cycle, the electrode circuit being suitably damped to prevent oscillations. For instance, a field of 10^5 volts/cm. applied by an electrode 300 cm. in length would produce a deflection of about 5 cm. in particles of enegy 10^9 ev. Such an electric deflection might be combined with a magnetic shielding channel such as that described by Skaggs and others (1946) for use with the betatron, especially as the long time of acceleration would permit the mechanical insertion of such a channel after the orbits had settled down, thus avoiding the possible damaging disturbance of the orbits due to distortion of the magnetic field during the injection period, when the magnetic field is small.

§8. THE VACUUM CHAMBER

Acceleration systems which employ a varying magnetic field necessitate use of a vacuum chamber, the walls of which cannot carry appreciable eddy currents which would upset the phase and shape of the field. The so-called "dough-nut" in the betatron is made of glass or ceramic and is coated on the inner surface with a thin shielding layer of silver or other metal, but this is not a practical solution for a large synchrotron. Figure 4 gives a schematic view of the chamber proposed for the apparatus in Birmingham. Models are under construction and the final chamber may differ in detail from this.

The acceleration space is formed of corrugated strips of stainless steel, the widths of which are small enough to reduce the effects of eddy currents to small proportions. The strips, which are radially disposed, are fixed rigidly to the outer octagonal section, also of stainless steel, and are not in electrical contact except at this junction. The whole is rendered air-tight by stretching a sheet of non-porous rubber over the strips and clamping this tightly to the octagon. The rubber is shielded completely from the beam by the interlocking corrugations, and since these are short compared with the wavelength of the radio-frequency accelerating field, the rubber is not subject to appreciable high frequency fields. The flat faces of the octagon are closed with plates which carry the pump manifolds, exit port, the insulated leads to the accelerating electrode, the source of protons, deflecting electrode etc. The chamber is made in eight sections bolted together with suitable rubber gaskets in each joint, and exhausted by six oil-diffusion pumps each 15 inches in diameter, in the manifolds of each of which is a liquid air trap.

The accelerating electrode is laminated to eliminate the effect of eddy currents. The advantage of this construction for the chamber is that it gives easy access to the interior and allows of modifications to the electrode system etc. without removing the vacuum system from the magnet. Such flexibility is important in equipment which is experimental in design, and where considerable modification may be required as a result of practical experience.

§ 9. SYNCHROTRONS FOR HIGHER ENERGIES

The synchrotron can be extended to higher energies by increasing the radius of the orbit and scaling up the other dimensions, keeping the same magnetic cycle. Figure 5 shows the way in which the cost, power demand and radius vary with the energy. It is seen that while equipment may be built for energies of 2–3 × 10^9 ev. it would be prohibitively expensive to construct a synchrotron for 10^{10} ev., at any rate in Great Britain. If higher energies are needed another method of acceleration must be used.

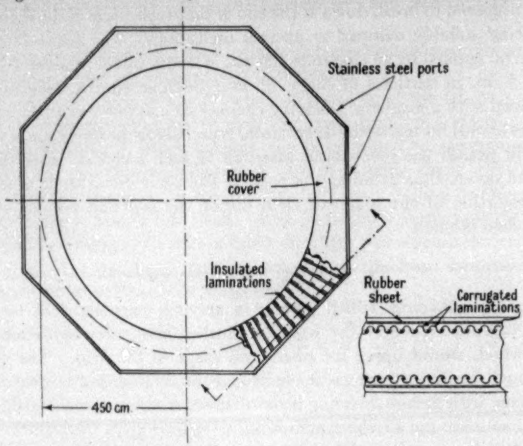


Figure 4. Schematic view of vacuum chamber.

§ 10. ACCELERATION OF ELECTRONS

It is impossible that energies as high as 10^9 ev. can be reached with electrons in a synchrotron of this type because of the excessive loss of energy as radiation by the particles due to their motion in a circle. This radiation loss can be calculated (Schwinger, 1945 and Schiff, 1946) and at 10^9 ev., with $\rho = 450$ cm., it is of the order of 20000 ev. per revolution. Thus to obtain electrons with energies comparable with those which can be reached with protons, the particles should be accelerated much more continuously, i.e. a large number of accelerating gaps would be needed with a voltage across each which is high compared with the nett rate of gain of energy, and driven at a correspondingly higher frequency.

It is difficult to estimate the maximum electron energy which could be obtained by operating the present equipment with a constan telectrode frequency of about 10 Mc., but it is probably in the region of 300–400 Me v.

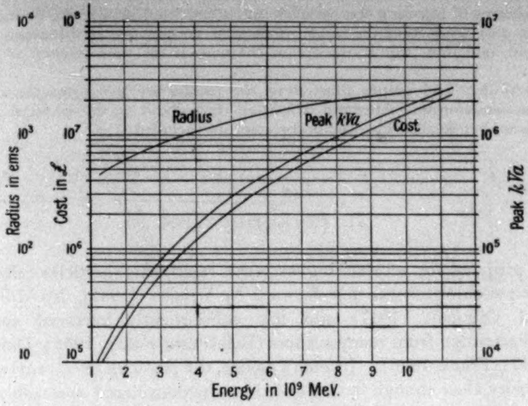


Figure 5. Variation of cost power demand and radius with particle energy.

REFERENCES

Gooden, Jensen and Symonds, 1947. *Proc. Phys. Soc.*, **59**, 677.
Goward and Barnes, 1946. *Nature, Lond.*, **158**, 413.
Kerst, 1941. *Phys. Rev.*, **60**, 47.
Kerst and Serber, 1941. *Phys. Rev.*, **60**, 53.
McMillan, 1945. *Phys. Rev.*, **68**, 143.
Peierls, 1946. *Nature, Lond.*, **158**, 773.
Schiff, 1946. *Rev. Sci. Instrum.*, **17**, 6.
Skaggs, Almy, Kerst and Lanze, 1946. *Phys. Rev.*, **70**, 95.
Veksler, 1945. *J. Phys. U.S.S.R.*, **9**, no. 3.
Wasserab. Unpublished.
Wideroe, 1928. *Arch. Electrotech.*, **21**, 387.

THEORY OF THE PROTON SYNCHROTRON

By J. S. GOODEN, H. H. JENSEN and J. L. SYMONDS

Read 20 *December* 1946 ; *MS. received* 21 *March* 1947

ABSTRACT. In the type of synchrotron for accelerating protons, the particle velocity, and consequently the accelerating radio-frequency, increase with increasing particle energy. In such a case the particle motion acquires properties which necessitate a careful control of some of the physical variables. In particular, it is found that within the stable limits of phase, non-relativistic particles, to a first approximation, possess undamped phase oscillations. The particles can be accelerated only so long as it is ensured that any factors affecting the phase oscillation amplitude are sufficiently small. It is necessary, therefore, to consider in some detail the physics of these oscillations, and in particular, of their damping.

It is found that there are no less than eight significant forces which can affect the behaviour of the phase oscillation amplitude. Four of these forces can be adjusted to some extent, the limitations being those of a practical nature. Thus it should be possible to accelerate protons in a synchrotron, if reasonable care is taken.

The problems of injecting the particles into a synchrotron working as such are considered. In this connection the radial oscillations accompanying the phase oscillations are described, and from this knowledge of the motion the time interval of injection is determined.

Numerical data and graphs illustrating the results are given for the case of the Birmingham synchrotron. Attention is focused throughout on the physical description of the motions, but detailed mathematical results are included.

---

§ 1. INTRODUCTION

General

THE proposal for accelerating extreme relativistic particles (electrons) by the synchrotron was put forward by Veksler (1945), McMillan (1945) and Oliphant (1947) and has subsequently received considerable theoretical attention from many authors (Bohm and Foldy, 1946; Dennison and Berlin, 1946; Frank, 1946). In such a device, the particles (electrons) are moving with a velocity close enough to that of light to be considered as sensibly constant. In this case the properties of the particle motion make it unnecessary to control stringently most of the physical variables of the system. In particular, it is found that for a large range of phases, the phase oscillation amplitudes decrease moderately rapidly with increasing particle energy, and thus the factors affecting the phase oscillation amplitude do not have to be carefully controlled. It is convenient to use the principle of the betatron to accelerate the electrons to velocities sufficiently close to that of light, before the synchrotron operation is commenced (Pollock, 1946). Then initial injection problems are identical with those of the betatron (Kerst and Serber, 1941).

In the type of synchrotron suggested by Oliphant and McMillan for accelerating protons, the particle velocity, and consequently the accelerating radio frequency, increase with particle energy. In such a case the particle motion acquires properties which necessitate a much more careful control of some of the physical variables. In particular, it is found that within the stable limits of phase (see below) the particles, to a first approximation, undergo undamped phase oscillations. If the phase oscillation amplitude is allowed to increase, the phases eventually reach the unstable region and the particles are lost. The particles can be accelerated only so long as any factors affecting the phase oscillation amplitude are sufficiently small. It is thus imperative to have a more thorough understanding of the physics of the phase oscillations and in particular of factors causing variations in the phase oscillation amplitude.

Initial acceleration by a betraton action is not practicable with protons. It is not possible to inject protons with energies high enough for them to be accelerated only in the relativistic range of velocities where they are inherently stable. It therefore becomes necessary to investigate the motions of particles following their injection. In what follows these problems are considered for the Birmingham synchrotron (Oliphant, Gooden and Hide, 1947).

Basic description of phase oscillations

The phase oscillations occurring in a synchrotron accelerating extreme-relativistic particles have been described by Veksler (1945) and McMillan (1945). Since the physical properties of these oscillations constitute the most part of what follows, it will be convenient to mention here in detail some of the essential features of such an oscillation. The variation of the amplitude of the oscillations and other refinements are treated in later sections.

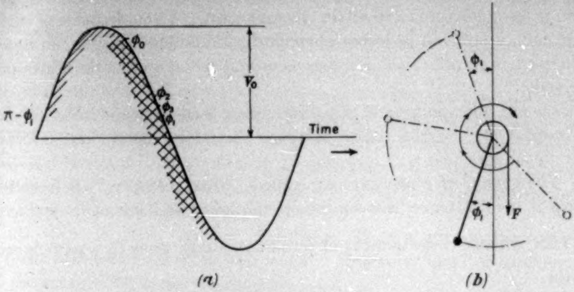


Figure 1. (*a*) Accelerating voltage amplitude to illustrate phase oscillations. (*b*) Pendulum analogue.

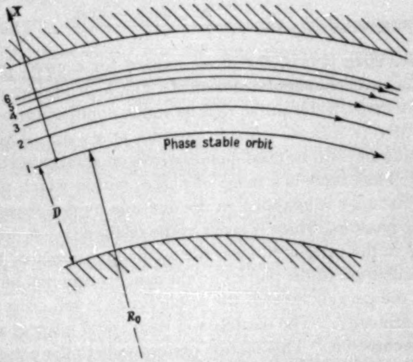


Figure 2. Illustrating the progressive increase in radius of a particle undergoing phase oscillations.

The basic action of the phase oscillations can be seen by reference to figures 1 and 2. It can be shown that the number of accelerating gaps is of no significance to the argument which follows and so, for simplicity, a one-gap system is assumed throughout. Consider firstly the case of extreme-relativistic particles, and assume the particles travel in circular orbits. The voltage amplitude, applied across the R.F. accelerating gap, is made greater than the energy (in volts) required to be added to a particle per revolution, in order to maintain it on the central orbit. The R.F. is so chosen that a particle moving on this orbit will always arrive at the gap at a certain constant R.F. phase. This phase is ϕ_1, so that the energy to be added per revolution to maintain the particle on the central orbit is $eV_0 \sin \phi_1$. This central orbit is called the stable orbit. A particle moving instantaneously on the stable orbit but arriving at the gap at a phase ϕ_0, say, receives extra energy and will thus increase its radius. This means that now the particle will lose in phase each revolution because of its larger circular path. It will therefore change its phase on next arriving at the gap in the direction of ϕ_1, gain excess energy and further increase its radius. This process will continue until the particle reaches an orbital radius and a phase such that the energy gain per rev. is just sufficient to maintain it on its orbit. Then, because of its larger circular path, it still will be losing phase and, gaining less energy than necessary to maintain it in its orbit, will decrease its orbital radius so that the whole process is reversed.

This action, when continued, sets up the phase oscillation and its accompanying radial oscillation. Several useful relations should be stated here. (Notation is given in the appendix.)

(1) The change of phase per rev. (called " phase velocity ") is directly proportional to the difference between the orbital radius and that of the stable orbit, i.e. $\frac{\partial \phi}{\partial q} \propto \delta R$, where q is the number of revolutions undergone by a particle starting from rest.

(2) The change in radius per revolution of the particle (rate of change of radius) is directly proportional to the ratio of the excess energy received by the particle per rev. to the total energy of the particle, i.e.

$$\frac{\partial R}{\partial q} \propto \frac{eV_0(\sin\phi - \sin\phi_1)}{E}.$$

(3) It follows that the acceleration of phase $\left(\frac{\partial^2 \phi}{\partial q^2}\right)$ is proportional to $\frac{\partial R}{\partial q}$ and consequently $\frac{\partial^2 \phi}{\partial q^2}$ varies as $\frac{eV_0(\sin\phi - \sin\phi_1)}{E}$. Thus, as was first pointed out by McMillan (1945), the phase motion for a given particle energy is identical in form with the angular motion of a simple pendulum under the additional influence of a constant torque so that its stable equilibrium position is ϕ_1 (see Figure 1 b). There is also an unstable position at $\pi - \phi_1$. This analogy is of considerable value and will be used quantitatively in another section.

Thus there is a range of phases within which the phase can oscillate stably. This range is bounded on the one side by the phase $(\pi - \phi_1)$ and on the other by the phase ϕ_2, which is given by the relation $(\pi - \phi_1 + \phi_2)\sin\phi_1 + \cos\phi_2 = 1 - \cos\phi_1$. As in the pendulum case, the oscillations about ϕ_1 and the corresponding radial oscillations of the particle in the synchrotron will be asymmetrical. If the phase of the particle exceeds the stable limits it will cease to oscillate and will increase continually. The particle will then spiral inwards and hit the inside wall of the synchrotron. This motion corresponds to the pendulum swinging continuously in a circle under the action of the constant torque when the bob is placed outside the stable limits.

In what follows reference will be made to the " phase restoring force " which, on the above argument (see point (3) above), will be proportional to

$$\frac{eV_0(\sin\phi - \sin\phi_1)}{E}.$$

All arguments concerning the conservation of energy of a pendulum have their valid analogy in the phase oscillation case.

When the particle is non-relativistic, and the magnetic field is uniform, the cyclotron conditions hold and there can be no phase stability. Furthermore, because of the increased orbital radius of a particle which receives excess energy, the particle will enclose a larger amount of changing magnetic flux, be further accelerated on this account and thus eventually hit the synchrotron walls.

If now the magnetic field is made to decrease radially outwards (a condition which is necessary to ensure vertical stability) the particle, on gaining excess energy (and velocity), will increase its radius more than when in a uniform magnetic field. The extra path thus introduced will cause the particle to change its phase as it rotates, just as in the extreme-relativistic case. Thus phase stability is introduced. The effect of the induction forces is now to increase the amplitude of the radial oscillations accompanying the phase oscillations, but they cannot prevent the oscillations from occurring.

Thus there is little essential difference, in principle, between the non-relativistic and extreme-relativistic cases. However, large differences occur in the damping of the phase oscillations.

§2. INJECTION

Particle motions

A particle injected into the synchrotron has, in general, two resulting radial oscillations. One is identical with the radial component of the oscillation occurring in a betatron and is described by Kerst and Serber (1941). This will be termed the " injection oscillation". The other radial oscillation is that accompanying the phase oscillation and has just been described.

In order to obtain a picture of the injection process, consider particles of uniform energy being directed continuously into the synchrotron. Neglect at first the injection oscillations. Then every particle will experience the phase and radial oscillations as described in the previous section, provided it enters the gap during the stable region of R.F. phase. Of these, only particles with maximum radial oscillation amplitudes less than the half width of the accelerating chamber will continue their motion and be accelerated. At a certain time, during a certain R.F. cycle, the particles entering will have their instantaneous orbit coinciding with the central stable orbit. For convenience this R.F. cycle is called the *zero*th R.F. cycle. Particles arriving during earlier or later R.F. cycles will have instantaneous orbits greater or less, respectively, than the stable orbit. These R.F. cycles are called the $-u$th and $+u$th respectively.

Consider particles entering during the stable phase range of the uth R.F. cycle. They will form a long bunch slightly inclined to the instantaneous orbit of the particle entering at the phase ϕ_1 (see figure 3). Because all the particles are moving on an orbit smaller in radius than the stable orbit, they will all have the same initial phase velocity and their phases will move up the voltage curve of figure 1. The particle arriving at phase ϕ_1 will just begin to gain excess energy and will therefore begin to increase its orbital radius. Those particles in front of the bunch, receiving much more than the stable energy, will increase their orbital radii rapidly, while conversely those at the back of the bunch will decrease their orbital radii, but more slowly. Consider the point 0 which moves along the stable orbit with the phase velocity of the R.F. Then the bunch of particles, as a whole, will rotate around 0, roughly remaining tangent to the curve traced out by the particle entering at phase ϕ_1. Successive stages of this motion are shown in figure 3. It is readily seen what determines the length of bunch which can be accelerated, and consequently the accepting phase range for the *u*th R.F. cycle. This is modified by the injection oscillations. The curve traced out by any particle is in general not simple, but for small amplitudes and for a rectangular reference system it becomes an ellipse.

If now the injection oscillations are also considered, the resulting motion of a particle injected during the *u*th R.F. cycle will be as shown in figure 4. The length of bunch accepted during a given R.F. cycle is given by the intercept of the instantaneous orbit of the particle arriving at phase ϕ_1 on the mean curve traced out by the particle which just misses the walls of the chamber. This determines the phase range of acceptance. The number of R.F. cycles accepting particles is obtained from this picture by determining when the instantaneous orbit added to the injection oscillation amplitude just touches the synchrotron chamber walls (or unstable region of magnetic field). Because of the random phases of the various oscillations among the particles from all R.F. cycles, the bunches will mix together to give a large resultant bunch of roughly uniform density and of a shape given by the envelope of the particle motion shown in figure 4.

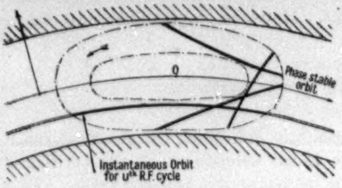


Figure 3. Successive stages of the motion of a bunch of particles entering during one R.F. cycle.

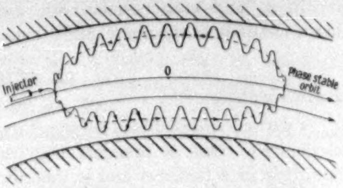


Figure 4. Motion of a particle injected into the synchrotron relative to a stable particle.

For a non-relativistic particle (as will be shown later) this big bunch will decrease in width as (kinetic energy)$^{-1/2}$ but will not decrease in length. For the Birmingham proton synchrotron the decrease in length, due to the relativistic phase damping introduced in the latter half of the acceleration, is only about four times, whereas the width of the bunch is reduced to a few millimetres. For an extreme-relativistic electron accelerator, the decrease in length is proportional to (total energy)$^{-1/4}$ and the width decreases more rapidly than (total energy)$^{-5/4}$.

Summing the number of R.F. cycles and the time intervals per R.F. cycle over which the particles are accepted, gives the total effective time interval, *T*, during which particles entering the synchrotron are eventually accelerated. For the type of injection system proposed for the Birmingham synchrotron

(Oliphant, Gooden and Hide, 1947), this time interval will be of paramount importance in determining the number of particles which will be accelerated.

Factors affecting the time interval of injection

It has been seen that the effective time interval of injection is made up of two parts: (*a*) the number of R.F. cycles accepting particles, (*b*) the phase ranges for each R.F. cycle over which particles can be accepted. These permissible phase ranges are limited by two conditions. The first limitation is that the R.F. phase at which a particle arrives at the gap must always be inside the stable phase range discussed in section 1. The second limitation is that the radial oscillation arising from the phase oscillation must always lie inside the synchrotron. These two limitations, in general, determine the permissible phase range for each R.F. cycle. It is evident that this phase range will be a maximum when the two ranges are made identical by adjusting the voltage amplitude on the accelerating electrodes (see below).

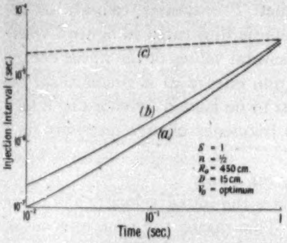


Figure 5. Variation of injection interval with time of rise of magnetic field to 15000 gauss :

(*a*) For injection energy $\epsilon_0 = 0{\cdot}3$ Me v.
(*b*) $\epsilon_0 = 1$ Me v.
(*c*) See text.

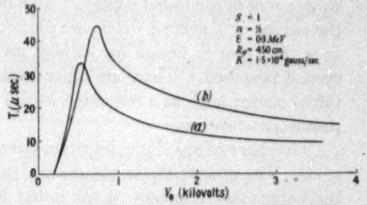


Figure 6. Variation of injection interval, T, with accelerating voltage amplitude, V_0 :

(*a*) $D = 15$ cm.
(*b*) $D = 20$ cm.

Both the number of accepting R.F. cycles and the accepting time intervals per R.F. cycle are influenced by a large number of physical variables. The most important of these are the rate of change of magnetic field, the initial energy, the voltage amplitude applied to the accelerating electrode and the width and depth of the accelerating chamber. The influence of such variables in relation to design is now considered in detail.

Rate of rise of magnetic field

Figure 5 illustrates the advantage to be gained by decreasing the rate of rise of the magnetic field. The conditions are those of optimum voltage amplitude, so that the increase in T obtained represents an increase of the optimum values. The corresponding voltages are also shown. The reasons for this behaviour of the effective time interval T are two-fold. Firstly, as $\partial B/\partial t$ is reduced, the time taken for the field to change the amount necessary to bring the instantaneous orbits from one side of the accelerating chamber to the other is increased proportionately to $(\partial B/\partial t)^{-1}$, and the time interval accordingly is increased. Secondly, as $\partial B/\partial t$ is decreased, the energy to be added per revolution is decreased. This means that the voltage amplitude must also be reduced in order to regain the optimum phase range. Because of this reduced amplitude, the excess energy received by a particle at a given phase will be less than before. Thus particles arriving at a given phase (same initial energy) will not increase their radii as much as before. This means that particles can be accepted over a larger phase range per R.F. cycle. The voltage is then reduced sufficiently to optimize T, thus making the increased phase range available. Obviously there is a limit to the increase gained in this way as ϕ_1 approaches zero. The dotted curve in figure 5 shows the expected mean current determined on this basis for a continuously oscillating magnetic field as a function of the oscillating frequency.

Initial energy

The injection energy (ϵ_0) influences the effective time interval of injection T in three ways, the total effect being that T increases roughly as the square root of the initial energy. This behaviour is illustrated in figure 5 where the curves of T against $\partial B/\partial t$ are plotted for several values of the initial energy. It is thus best to design for as high an injection energy as is practical for the system proposed. There are other advantages to be gained by choosing a high initial energy such as a reduction in the radio frequency change necessary for a proton synchrotron.

The three effects of the injection energy, ϵ_0, on T are:

(1) The greater the initial energy, the greater the initial velocity, and hence the higher the frequency of the initial R.F. This means the time intervals of all given phase ranges are reduced in the proportion $(\epsilon_0)^{-\frac{1}{2}}$.

(2) As ϵ_0 increases, so does the magnetic field required to give the same orbital radius. A greater absolute change in the magnetic field is needed to move the instantaneous orbit of the particle from one side of the synchrotron chamber to the other. The rate of change of magnetic field is the same, so that a larger time will elapse for this process $(\propto \epsilon_0^{\frac{1}{2}})$ and consequently the number of accepting cycles is increased $(\propto \epsilon_0^{\frac{1}{2}})$ and this will make the total effective time interval "T" greater.

(3) From a similar argument it follows that the particles, since they undergo phase oscillations and the related radial oscillations, must now gain a greater amount of excess energy to have a radial oscillation amplitude of the same size as before. This increases the accepting phase ranges proportionately to $(\epsilon_0)^{1/2}$.

Combination of these three effects produces an increase in the effective time interval of injection proportional to ϵ_0, for small phase oscillation amplitudes.

Voltage amplitude

The variation expected in T as a result of varying the voltage amplitude applied to the accelerating electrode is shown in Figure 6. The existence of an optimum voltage has already been mentioned and is chosen so that the two limits of accepting phase range coincide. If the voltage amplitude decreases from this optimum, the stable phase ϕ_1 will increase and, the upper phase limit being $\pi - \phi_1$, the accepting half range is $\pi - 2\phi_1$ and will steadily decrease to zero. If the voltage amplitude is increased from the optimum, then ϕ_1 decreases and the upper limit, set by the radial oscillations, decreases rapidly. Then the accepting phase range goes asymptotically to zero for large voltage amplitudes.

Effective width of the accelerating chamber

By the effective width is meant the actual width minus twice the amplitude of any radial oscillations other than those already mentioned. Such oscillations arise from the circular irregularities in the magnetic field coming from eddy currents, inhomogeneities in the iron or construction. The effect of the effective width on T is illustrated in figure 6. It is seen that considerable advantage is to be gained by increasing this width.

The reason for this increase is not hard to see. A larger accelerating space means more R.F. cycles can accept particles, the number of cycles being proportional to the width of the chamber. Furthermore, the particles can now have larger radial oscillation amplitudes and this permits larger phase ranges. In the case of optimum voltage amplitudes this allows the voltage optimum to increase, thereby decreasing ϕ_1 and increasing the stable limits of phase. The increase in phase allowed on this account will involve the injection oscillations, whose amplitudes will be different for each instantaneous orbit. Thus the position of the injector will modify the effective increase in phase range gained in this way.

Depth of accelerating chamber

Besides the advantage of being able to tolerate larger vertical disturbances to the particles, increasing the accelerating chamber depth provides another advantage in the case of the Birmingham synchrotron (Oliphant, Gooden, Hide, 1947). Here it is proposed that the stable orbital plane be lowered continuously during the period of injection. The amount the orbital plane has been lowered before the particles return to the immediate neighbourhood of the injector determines the effective thickness of the proton beam. Thus a greater depth of the accelerating space allows the orbital plane to be lowered by a proportionately greater amount. The depth should consequently be as large as can be tolerated on other grounds. The depth affects very critically the energy stored in the magnetic field, other dimensions remaining constant as both the volume of space is increased and the magnetic field reduced for the same ampère-turns. It does not affect the beam-current as strongly as does the width, and so a ratio of depth to width of about 1 : 2 seems a good choice.

Table 1 illustrates what effective thicknesses of injected proton beam can be obtained for different depths of accelerating chamber for the Birmingham synchrotron.

Table 1

Depth of chamber (cm.)	7·5	10	15
Beam thickness (mm.)	2	3	4·5

Quantitative results

The general quantitative information concerning the injection process can be obtained without solving the phase equation at all. Remembering that there is a relation between the phase velocity and the difference between the orbital radius of the particle and the stable orbit radius, the analogy of the conservation of energy of the pendulum bob can be employed. The magnetic field is here assumed to rise linearly in time. A departure from this condition makes little difference to the magnitude of the results but considerably complicates the form they take. Then the maximum phase velocity and consequently the maximum orbital radius can be determined for particles entering at different phases and times.

The first relation is, in fact:

$$\frac{\partial\phi}{\partial q}=\frac{2\pi np}{R_0}(R_0-R) \qquad \ldots\ldots(1)$$

using the notation given in the appendix. If ϕ_0^u is the limit (upper or lower) of the acceptable phase range of the uth R.F. cycle, then, using the analogy just referred to, ϕ_0^u is given by

$$|\cos\phi_1-\cos\phi_0^u-(\phi_0^u-\phi_1)\sin\phi_1|^{1/2} \leqslant \frac{\sqrt{2}\pi np(q_0+u)^{1/2}}{R_0A^{1/2}}\{[D-|x_1-x_0^u|]^2-(x_0^u)^2\}^{1/2}, \qquad \ldots\ldots(2)$$

where

$$A=\frac{V_0\mathbf{c}np}{2KR_0^2(1-n)}. \qquad \ldots\ldots(3)$$

From this, by graphical methods of integration, the total effective time interval of injection T can be obtained. The number of accepting R.F. cycles is given by

$$N=\frac{2D(1-n)\epsilon_0 p}{\mathbf{e}R_0V_0\sin\phi_1}. \qquad \ldots\ldots(4)$$

In the case when the phase lies on the approximately straight portion of the R.F. voltage curve, T can be integrated and becomes:

$$T=\frac{\dfrac{(3\cdot3)\mathbf{c}}{\mathbf{e}}\sqrt{\dfrac{m}{\mathbf{e}}}[n^{1/2}(1-n)^{3/2}]\left[\dfrac{D(D^2-x_1^2)^{1/2}}{R_0^3}\right]}{K(V_0\cos\phi_1)^{1/2}}\epsilon_0 p^{1/2}. \qquad \ldots\ldots(5)$$

For the optimum choice of V_0, $\phi_0^0=\pi-\phi_1$, and since

$$V_0=\frac{2\pi R_0^2K}{\mathbf{c}\sin\phi_1}, \qquad \ldots\ldots(6)$$

in this case, the optimum value of ϕ_1 is given by

$$|2\cos\phi_1-(\pi-2\phi_1)\sin\phi_1|\operatorname{cosec}\phi_1 =\left\{\frac{np(D-x_1)}{R_0}\right\}^2\left\{\frac{(1-n)\epsilon_0\mathbf{c}}{Kn\mathbf{e}p}\right\}. \qquad \ldots\ldots(7)$$

For the case when $\phi-\phi_1\ll\phi_1$ and $\phi-\phi_1\simeq\sin(\phi-\phi_1)$, we have

$$(\pi-2\phi_1)\left\{\tan\left(\frac{\pi-2\phi_1}{2}\right)\right\}^{1/2}=\frac{n(D-x_1)}{R_0^2}\left\{\frac{2(1-n)\epsilon_0\mathbf{c}p}{Kn\mathbf{e}}\right\}^{1/2}, \qquad \ldots\ldots(8)$$

which determines the optimum value of ϕ_1.

§ 3. PHASE OSCILLATION AMPLITUDE BEHAVIOUR

There are no less than eight significant forces which can affect the behaviour of the phase oscillation amplitude. In order to introduce these forces in the most convenient way, those which cannot be externally varied are considered, firstly by discussing the simpler case of extreme-relativistic particles and then by extension to the case of non-relativistic particles. Intermediate regions follow immediately. There are then left four independent adjustable forces which are considered in turn. These latter can be employed to give a useful range of permissible variation of those factors which cause an increase in the phase amplitude. Finally, the effect of these various forces is considered quantitatively and curves given to enable the value of the forces to be assessed.

For an extreme-relativistic particle (i.e. sensibly constant velocity) accelerated in a synchrotron whose magnetic field increases linearly with time and whose accelerating voltage amplitude and R.F. are maintained constant, there is one damping force together with two opposing (anti-damping) forces.

The damping force can be explained in the following way. In the description of the phase and radial oscillations of section 1, it was shown that the particle reached a maximum or minimum orbital radius when the energy it received per revolution was just sufficient to maintain it on this circular orbit. This orbit marked the reversal of the phase force, as thereafter the phase acceleration changes sign, since the change in phase per revolution begins to decrease. When a particle is moving on a larger radius than that of the stable orbit, it will require a greater energy increase per revolution to maintain it on that larger radius. Thus the particle will reach its maximum or minimum radius when its phase reaches ϕ_2 (see figure 1), and it is at this point that the phase restoring force is reversed. This is the reason for the damping and the process is quite analogous to the damping of an oscillator in a viscous medium. An important observation to note here is that the rate of damping will be different for phase oscillations of different amplitudes. So long as the phases remain on the approximately straight portion of the R.F. curve, the phase velocities, and therefore the damping force, will be in proportion to the phase amplitude (cf. pendulum). Phase oscillations with amplitudes extending beyond this approximately straight portion will have their maximum phase velocities less than proportional to the amplitudes, and thus the damping rate will be less. This behaviour applies to all the forces affecting the phase amplitude.

A force which opposes this damping force is the one arising from the electric field of induction (betatron force). For the Birmingham synchrotron with a magnet yoke on the inside of the air gap, the changing return flux gives a decelerating force acting on the particle. The changing flux in the air gap gives an accelerating force, but since only part of this flux is ever enclosed by the particle orbit, the nett result is a decelerating force which decreases with increasing orbital radius. Consequently a particle increasing its orbital radius will require less energy per revolution from the R.F. accelerating field to maintain it on its radius than on the previous argument. The phase force is then reversed, not at ϕ_2, but at an intermediate value, say at ϕ_3 in figure 1. This opposing force is always a certain fraction less than unity of the damping force. The results of the above discussion are not affected by the position of the return flux from the air gap.

Another factor which increases the phase amplitude is the increase in energy of the particle. The action is not anti-damping in the sense of adding " phase energy ", but is analogous to " conserving phase energy ". As the energy E of the particles increases, the change in radius experienced by the particle on receiving a given excess energy of amount $\delta\epsilon$ decreases in proportion to $1/E$. But it is this change in radius per revolution which determines the " phase restoring force " and thus this restoring force is reduced ($\propto 1/E$). The frequency of the phase oscillation is reduced thereby ($\propto E^{-1/2}$) and to conserve "phase energy" the amplitude of the phase oscillation must increase ($\propto E^{1/4}$).*

These three major factors give a resultant damping to oscillations occurring on the straight portion of the R.F. curve, proportional to $E^{-1/4}$ (Veksler (1945), McMillan (1945), Bohm and Foldy (1946), Dennison and Berlin (1946), Frank (1946)). For larger amplitudes this rate of damping will decrease.

When a non-relativistic particle is accelerated, the R.F. must increase in step with the mean particle velocity. Although the same arguments as above also apply to the non-relativistic case, this changing R.F. introduces another force which still further reduces the rate of damping. Since the rate of change of R.F. is chosen to keep in step with a particle maintained on the stable orbit, it follows that the frequency will not increase quickly enough for a particle on a radius greater than the stable orbit. Hence the phase of such a particle will experience an extra acceleration, or in other words, the change in phase per revolution of this particle will increase every revolution. This means that as a particle increases its radius (decreases its phase) it also experiences an increase in its phase acceleration, or phase force, on this account. This action is just the opposite to damping and so the resultant damping due to the previous three forces is still further reduced. It is shown quantitatively below that the resultant of all these forces is to give no damping at all. Since their behaviour is dependent on the maximum phase velocity, the variations in their effects on oscillations of different amplitudes will be the same and so the damping will be zero for all oscillation amplitudes.

In the actual case of acceleration of protons to energies of about 1000 Mev., the finite relativistic effects cause a small damping action during the latter part of the acceleration.

Adjustable damping factors

Four methods exist whereby the behaviour of the phase oscillation amplitude can be adjusted. These are :—(*a*) changing the rate of change of R.F., (*b*) varying the voltage amplitude during acceleration, (*c*) varying the way in which the magnetic field increases with time, (*d*) shaping the faces of the accelerating electrode. Each of these will now be considered qualitatively and quantitatively.

Further variation in R.F.

From the argument given in the previous paragraph it follows that if the R.F. continually increases its rate of change more rapidly than to maintain a particle on any fixed radius, there will be an increase in the phase oscillation amplitude—and conversely. It also follows that this effect arising from the factor $d^2\nu/dt^2$ will depend on the change in radius of a particle for a given energy change, and this in turn will depend on the way in which the magnetic field falls off radially (i.e. on n). For instance, if $n=0$, then any finite continuous change to the R.F. will cause infinite anti-damping in the non-relativistic case, and the larger n, the less will be this effect. On this basis, it is best to design for a large n and then a large tolerance on frequency variation can be allowed.

---

* The quantitative results given here in brackets refer accurately only to small oscillation amplitudes.

Figure 7 shows quantitatively the continuous increase in R.F. in the case of the Birmingham synchrotron, giving a loss of about 10 % of the particles by anti-damping action. This increase in R.F. is plotted as a tolerance against the value of the magnetic field index n for different ways of increasing the magnetic field in time (different values of s).

The effect of this variable damping force is given later by equation (15). It should be noted that the rates of damping or anti-damping always decrease with increase of oscillation amplitudes, for reasons given above.

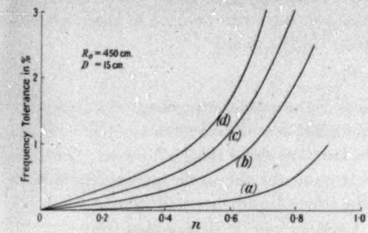


Figure 7. Variation of frequency tolerance due to anti-damping with magnetic field index n: (*a*) $s=0{\cdot}8$, (*b*) $s=1{\cdot}0$, (*c*) $s=1{\cdot}2$, (*d*) $s=1{\cdot}5$. 10 % particle loss allowed.

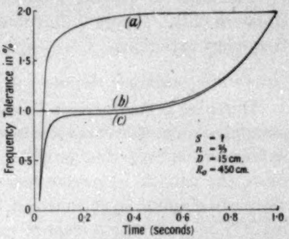


Figure 8. Variation of the frequency tolerance with time, allowing 10% loss of particles. Tolerance: (*a*) for orbital shift, (*b*) for anti-damping, (*c*) combined.

Variation of voltage amplitude

If the voltage amplitude is increased, then a particle arriving at an accelerating gap at a given phase will receive a larger excess energy, $\delta\epsilon$, than previously. In fact, $\delta\epsilon$ is proportional to V_0, the voltage amplitude, and so the rate of increase of radius is proportional to V_0. Thus by increasing the voltage amplitude during the acceleration, a particle will progressively increase its rate of increase of radius, i.e. its phase acceleration or phase restoring force ($\propto V_0$). This means that the frequency of phase oscillation is increased proportionately to $V_0^{1/2}$ and, to conserve "phase energy", the amplitude of the phase oscillation must decrease as $V_0^{-1/4}$.* Although this gives a method whereby the phase oscillations can be damped, it also means that the final width of the proton beam is larger, and this may interfere with extraction. The final width can be obtained from the relation between the radial oscillation and phase oscillation (equation 1). To obtain a reasonable effect, a large increase in voltage is required and this implies a very large increase in R.F. power.

Manner in which the magnetic field increases with time

Consider a cycle of the phase oscillations and let the magnetic field in the air gap increase as $B=Kt^s$, where t is the time and K a constant. Let $s>1$ for the present argument; this means that the energy required to be added per revolution in order to maintain a particle in a given orbit (R_0) must increase as t^{s-1}. To maintain a particle at some constant stable phase will therefore necessitate an accelerating voltage amplitude, increasing as t^{s-1}. This increase will be finite over any one phase oscillation cycle and so the total excess energy gained by a particle undergoing the phase oscillation will be greater than in the case of constant accelerating voltage. Since the percentage increase in the magnetic field during this period is small, the particle will increase its radius more quickly than when $s=1$. This means that it reaches its maximum radius, and hence the turning point of the phase force, at a greater phase difference from ϕ_1 than previously. Hence increased damping results. The converse holds when $s<1$, which is the case of a sinusoidally increasing magnetic field.

---

* See previous footnote.

The effect of varying s on the rate of damping is given quantitatively by equation (12). Figure 7 illustrates how varying s can be used to allow tolerable frequency variations.

Accelerating-electrode shaping

Damping is also increased if the faces of the accelerating electrodes are sloped, parallel to each other, at an angle to the radius vector, in such a way that a particle moving on a larger radius will arrive later in phase than otherwise. This will cause the particle to receive less acceleration for the same radius than without electrode shaping, an effect which, on the above arguments, will increase the damping. The quantitative value of such a system can be determined from equation (17). It is not certain how the inevitable radial disturbances thus introduced will affect the motion, but the method is an easy one with which to experiment.

Quantitative analysis

(1) *Non-relativistic case.*—Using the notation in the appendix, the phase equation for the non-relativistic particle can be written as:

$$\begin{aligned}
&\frac{\partial^2\phi}{\partial q^2} + \Bigg\{\frac{s}{s+1}(1-f(t))\left[1 + \frac{(\phi-\phi_1)}{2\pi pq} + \frac{1}{q}\int_0^q f(t)dq\right]\frac{1}{q} \\
&\quad - \left(\frac{2(1-n)}{n}\right)\frac{\partial f(t)}{\partial q}\Bigg\}\frac{\partial\phi}{\partial q} + \frac{V_0\mathbf{c}np}{(s+1)KR_0^{\,2}(1-n)}\Bigg\{1 + \frac{(\phi-\phi_1)}{2\pi pq} \\
&\quad + \frac{1}{q}\int_0^q f(t)dq + \frac{(1-n)}{2\pi np}\frac{\partial\phi}{\partial q}\Bigg\}\frac{\sin\phi}{q} = \frac{2\pi snp}{(s+1)(2-n)}\Bigg\{1 + \frac{(\phi-\phi_1)}{2\pi pq} \\
&\quad + \frac{1}{q}\int_0^q f(t)dq + \frac{(1-n)}{2\pi np}\frac{\partial\phi}{\partial q}\Bigg\}\Bigg\{1 - \frac{(2-n)}{n}f(t) + \left(\frac{R_1}{R_0}\right)^{2-n}\frac{1}{(1-n)}\Bigg\}\frac{1}{q} \\
&\quad - 2\pi p\frac{\partial f(t)}{\partial q}. \qquad \ldots\ldots(9)
\end{aligned}$$

Here the adiabatic theorem is employed, and $f(t)$ is the fractional increase in R.F. above that required to maintain the particle in the stable orbit. It is defined as

$$\text{R.F.} \equiv \nu = \frac{v_0 p}{2\pi R_0}(1+f(t)). \qquad \ldots\ldots(10)$$

Neglecting $f(t)$ for the moment, the equation becomes, putting $q=x^2$,

$$\frac{\partial^2\phi}{\partial x^2} + \frac{1}{x}\frac{(s-1)}{(s+1)}\frac{\partial\phi}{\partial x} + \frac{4V_1\mathbf{c}np\sin\phi}{(s+1)KR_0^{\,2}(1-n)} = \frac{2\pi snp}{(s+1)(1-n)}, \qquad \ldots\ldots(11)$$

where $V_0 = V_1 t^{s-1}$ is the voltage amplitude.

It is readily seen that for $s=1$ the damping term disappears and the resulting equation is identical in form with that of an undamped simple pendulum with a constant torque. Thus for all amplitudes there is no damping in this case.

For small amplitudes when $\sin(\phi-\phi_1) \simeq \phi-\phi_1$ and $(\phi-\phi_1) \ll \phi_1$, the equation can be solved as a Bessel equation, the asymptotic solution being always valid. The solution is then, putting $\sin\phi_1 = \dfrac{2\pi s K R_0^2}{V_1 \mathbf{c}}$, $C = \dfrac{V_1 \mathbf{c} n p \cos\phi_1}{(s+1)KR_0^2(1-n)}$,

Non-relativistic:

$$\phi = \phi_1 + G q^{\frac{1-s}{4(1+s)}} \sin\{2(Cq)^{\frac{1}{2}} + \alpha\}, \qquad \text{......(12)}$$

where G and α are arbitrary constants.

(2) *Extreme-relativistic*.—For the extreme-relativistic case it can easily be shown that the solution is

$$\phi = \phi_1 + G q^{\frac{1-2s}{4}} \sin\{2(C_1 q)^{\frac{1}{2}} + \alpha\}, \qquad \text{......(13)}$$

where

$$C_1 = \frac{V_1 p \mathbf{c} \cos\phi_1}{K(1-n)R_0^2}.$$

(3) *General case*.—In the general case intermediate between the non- and extreme-relativistic cases, and when $s=1$, the damping term can be shown to be given by

$$\text{amplitude} \propto q^{-\frac{g}{4(2+g)}}, \qquad \text{......(14)}$$

where

$$g \equiv \frac{\text{kinetic energy}}{\boldsymbol{m}_0 \mathbf{c}^2}.$$

Damping effect of frequency variations follows from the general equation in the case that $(\phi-\phi_1) \ll \phi_1$ and $\sin(\phi-\phi_1) \simeq \phi-\phi_1$ by a Bessel solution. The phase damping is then found to be as q^A, where

$$A = f(t) - \frac{1}{q}\int_0^q f(t)\,dq + \frac{4(1-n)q}{n}\frac{\partial f(t)}{\partial q} - \frac{\phi-\phi_1}{2\pi p q}. \qquad \text{......(15)}$$

The damping effect of voltage amplitude increase has already been shown by qualitative argument to be given by the term $V^{-1/4}$, i.e. due to this effect in the absence of others the phase varies as

$$\phi - \phi_1 = \text{const. } V^{-1/4} \cos(Dq^{1/2}V^{1/2} + \alpha), \qquad \text{......(16)}$$

where D is a constant. This can be confirmed by direct solution in many cases.

The damping effect of variations of the magnetic field increase has already been given by equations (12) and (13).

The damping effect of electrode shaping is given by the relation q^{-B}, where

$$B = 1 - \frac{ps \cot\phi_1 \tan\theta}{(s+1)(1-n)}, \qquad \text{......(17)}$$

where θ is the angle between the radius vector and the electrode face.

§4. PERMITTED VARIATIONS OF THE R.F.

Phase damping

Since the R.F. varies by a large factor, it is important to know the effects on the particle motions of small deviations of this frequency from the required law.

It has already been shown that if the frequency increases by about 2 % too rapidly, then 10 % of the particles are lost because the phase amplitude increases. Figure 7 gives the data on this effect for the Birmingham synchrotron. Figure 8 shows how in a given case this R.F. tolerance is relaxed towards the latter part of the acceleration because of the finite damping due to relativistic effects.

Orbital disturbances

It is readily seen, on the basis of the description of the particle motions given in the section on injection, that if the R.F. does not correspond to a stable orbit in the centre of the chamber, then a smaller number of particles will be accepted. In fact, the effective half width of the accelerating chamber is the shortest distance between the stable orbit and either synchrotron wall. If the R.F. changes slowly enough the particles can follow a change in the stable orbit. It then follows from the equations (1) and (11) and the reasoning in the section on injection that, in order to move the stable orbit a fraction L of the chamber half-width, $f(t)$ must obey the relation

$$f(t) \leqslant \frac{npD}{R_0}\left\{1-(1-L)\left(\frac{q}{q_0}\right)^{-1/2}\right\}. \qquad \text{......(18)}$$

For this fraction L, it is safe to assume that not more than a fraction $1-(1-L)^2$ of the particles is thus lost. Figure 8 shows the tolerance permitted for the Birmingham synchrotron when 10 % of particles are lost in this way.

For rapid R.F. variations (random oscillations) the particles have difficulty in following the variations in the stable orbit and so the permitted frequency variations are increased. The general condition is given by

$$\left|-\frac{(1-n)}{2n}(\sin 2\phi_1)\frac{\partial f(t)}{\partial q}+\frac{\partial F(q)}{\partial q}+2\pi p f(t)\right| \leqslant \frac{2\pi npD}{R_0}\left\{1-(1-L)\left(\frac{q}{q_0}\right)^{-1/2}\right\}, \qquad \text{......(19)}$$

where $F(q)$ is a particular solution of the equation

$$\frac{\partial^2\phi}{\partial q^2}+\frac{s}{s+1}\frac{1}{q}\frac{\partial\phi}{\partial q}+\left\{\frac{V_1 c n p \cos\phi_1}{(s+1)KR_0^2(1-n)}\right\}\frac{\phi}{q}=-2\pi p\frac{\partial f(t)}{\partial q}.$$

Thus, for $f(t)=k\cos\omega q$, say, and for $\omega \gg \frac{1}{2q}$, it is found that $\frac{\partial F(q)}{\partial q} \simeq -2\pi p . f(t)$ and the inequality for these rapid variations becomes

$$k \leqslant \frac{2\pi n^2 p}{(1-n)(\omega\sin 2\phi_1)}\frac{D}{R}\left\{1-(1-L)\left(\frac{q}{q_0}\right)^{-1/2}\right\}. \qquad \text{.....(20)}$$

For the Birmingham synchrotron rapid variations in frequency can be very large indeed.

Combined effects

When these two effects of the R.F. are combined, the tolerance on the R.F., determined by allowing no more than 10 % of particles to be lost, will vary during the acceleration period. For the particular case of the Birmingham synchrotron, this variation in the tolerance is shown in figure 8. It is to be noted that the strict tolerance necessary at injection is relaxed by an order of magnitude in about 1/50 of the total acceleration time.

REFERENCES

Bohm and Foldy, 1946. *Phys. Rev.*, **70**, 249.
Dennison and Berlin, 1946. *Phys. Rev.*, **70**, 58.
Frank, 1946. *Phys. Rev.*, **70**, 174.
Kerst and Serber, 1941. *Phys. Rev.*, **60**, 53.
McMillan, 1945. *Phys. Rev.*, **68**, 143.
Oliphant, Gooden and Hide, 1947. *Proc. Phys. Soc.*, **59**, 677.
Pollock, 1946. *Phys. Rev.*, **69**, 125.
Veksler, 1945. *J. Phys. U.S.S.R.*, **9**, no. 3.

APPENDIX

Notation used :—

$\pi-\phi$ = phase of R.F. (see figure 1).
ϕ_1 = stable phase.
R = radius of particle orbit.
R_1 = max. radius of magnetic field.
$R_0=\frac{1}{2}$ (min. radius + max. radius).
D = half width of accelerating space.
ν = radio frequency (R.F.).
$f(t)$ = relative variation in R.F. from the required law (see equation (10)).
V_0 = voltage amplitude appearing on the accelerating electrodes.
ϵ = kinetic energy.
ϵ_0 = kinetic energy at injection (time $t=t_0$).
E = total energy.
q = number of revolutions undergone by a particle starting from rest in the synchrotron, i.e. starting at zero time ($t=0$).
p = ratio of R.F. to particle frequency of revolution and is a positive integer.
T = effective time interval over which particles entering the synchrotron are accepted for acceleration to the peak energy.
t = time.
t_0 = mean time of injection.
B = flux density in the air gap and varies as $B=Kt^s\,(R_0/R)^n$, where
K = a constant, determining the rate of rise of B,
s = a positive index determining the way in which the magnetic field increases with time,
n = a positive index, less than unity, determining the way the magnetic field varies with radius.
x_1 = the radial coordinate of the injector, measured positively outwards from R_0.
$x_0^u+R_0$ = the radius of the instantaneous orbit of a particle entering at the uth R.F. cycle.
u = the number of R.F. cycles following the one when the instantaneous orbit has a radius R_0.
v_0 = velocity of particle in the stable orbit.

THE HOLE THEORY OF DIFFUSION

By G. WYLLIE,
Bristol

MS. received 23 *December* 1946

ABSTRACT. We show that in a dilute substitutional solid solution of one metal in another the diffusion of the solute atoms is determined by the compound activation energy $\epsilon_{A}+\epsilon_{ABA}$ where ϵ_{A} is the energy necessary for the formation of a hole next to a dissolved atom and ϵ_{ABA} the energy of activation for the hole to make one jump round that atom, provided $\epsilon_{ABA}-\epsilon_{BB}\gg kT$, where ϵ_{BB} is the energy of activation for the hole to diffuse away from the dissolved atom.

This mechanism provides an explanation of the cases where diffusion of a foreign metal atom in a lattice has a lower activation energy than self-diffusion in the same lattice, gold in lead being a conspicuous example. However, the assumption of next-neighbour interactions made in the description in this paper does not correspond to the facts of metallic structure. The statistical argument is not invalidated by this, but the calculation of the actual energies involved becomes a very difficult problem in quantum mechanics which has not yet been solved.

Johnson (1939) has given a semi-quantitative treatment of diffusion in dilute metallic solid solutions by the mechanism of Schottky defects. He pointed out that it might be energetically possible for a solute atom and a hole (i.e. a vacant lattice point) to adhere, and to wander together, with a comparatively low activation energy, through the matrix before separating. Such a process could explain the observed fact that the activation energy for diffusion of atoms dissolved substitutionally in a metal lattice is frequently less than the activation energy for self-diffusion in the same lattice.

The object of this note is to make a more complete analysis of the problem for a rather simple case, which, however, is not too far removed from experimental conditions. We consider N_{A} atoms of type A and N_{B} of type B ($N_{A}\ll N_{B}$) arranged on a cubic close-packed lattice. Then the problem is to determine the diffusion coefficient of the A atoms through the lattice, the movement of each A atom being by a jump into a neighbouring vacant lattice position. In talking about the relation of an atom to its nearest neighbours, it is useful to take a unit cell which is not face-centred but edge-centred, derived from the ordinary face-centred cell by a translation of half its edge parallel to an edge (figure 1). Let the centre atom in the cell in figure 1 be surrounded by B atoms except for the point H, which is vacant. Then there are four B atoms which lie next to the centre atom and to the vacant point, so in order to jump into the vacant position that atom has to squeeze through a gate of the form shown in figure 2, where the atoms are represented as spheres of diameter equal to the distance between the centres of nearest neighbours. Evidently it is to be expected that a foreign atom, if of small radius, should require a lower activation energy for jumping into a neighbouring hole than one of the matrix atoms (A_{1} and B_{1}, figure 2). However, a B atom next to an

A atom, both being next to a hole, has to pass through an asymmetrical gate (A_2, figure 2), and may thus require a lower activation energy for the jump than a B atom not so situated.

We may calculate the probabilities of the different possible jumps, in terms of the energy changes involved, by the reaction rate theory developed by Eyring and others. This is done by supposing the system in an "activated state" to move in an arbitrary small length δ of the reaction coordinate about the maximum of potential energy. Then the probability of a transition per unit time is given by the product of the probability of occurrence of this activated state and the thermal velocity in the reaction coordinate, divided by δ.

If each atom is considered to vibrate in the average field due to the others, we can write down the partition function for an arbitrary configuration of the system (i.e. for arbitrary numbers of vacant lattice points and activated atoms on the way into the vacancies), taking the pressure to be constant and zero. Then the equilibrium state is given by the maximum of the partition function.

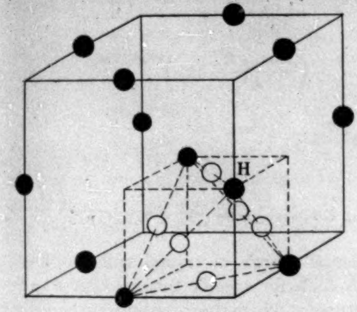


Figure 1.
Black circles indicate points of the metal lattice, white circles points of AH pair lattice.
N.B.—One black circle has been omitted for the sake of clarity.

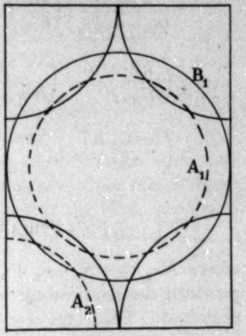


Figure 2.

Suppose n_B holes have only B atoms surrounding them, while n_A have an A atom as a nearest neighbour. We suppose also that N_A is so much less than N_B that the case of a hole having more than one A neighbour may be ignored. Of the n_B, let n_{BB} have an activated atom moving in. Of the n_A, let n_{AA} have an activated A atom moving in, n_{AB} an activated B atom not next to the A atom moving in, and n_{ABA} an activated B atom next to the A atom moving in. Let the energy required to remove a B atom from a position not next to an A atom, leaving a hole, be ϵ_B, and the energy to remove a B atom from a position next to an A atom be ϵ_A. Also let ϵ_{BB}, ϵ_{AA}, ϵ_{ABA}, ϵ_{AB} ($=\epsilon_{BB}$, if we consider only next-neighbour interactions) be the activation energies for the processes with the corresponding subscripts.

We neglect the changes in frequency of oscillation of atoms at the surface of holes, B atoms next an A atom and atoms between which an activated atom is just squeezing. These can be introduced if necessary in any particular case without altering the combinatory factor in the partition function, so merely modify the multiplying factors, not the exponential terms, in the final expressions for the rates. B atoms are taken to oscillate with frequency ν_B in all directions, A with ν_A in all directions, BB with ν_{BB} in both directions normal to the direction of the jump, AA with ν_{AA} in both directions, ABA with ν_{ABA1} in one direction and ν_{ABA2} in the other, AB with ν_{BB} in both. We assign arbitrary distances in the reaction coordinates $\delta_{BB,AA,ABA,AB(=BB)}$

The partition function F for the system is then given by

$$F=\left[\frac{N!\,(N-13N_A)!\,(12N_A)!\,12^{n_{BB}}\,.\,4^{n_{ABA}}\,.\,7^{n_{AB}}}{(N-N_A)!\,N_A!\,(12N_A-n_A)!\,n_{AA}!\,n_{ABA}!\,n_{AB}!\,n_{BB}!\times(n_A-n_{AA}-n_{ABA}-n_{AB})!\,(N-13N_A-n_B)!}\right.$$

$$\left.\times\frac{1}{(n_B-n_{BB})!}\right]\left(\frac{\boldsymbol{k}T}{\boldsymbol{h}\nu_B}\right)^{3(N_B-n_{BB}-n_{ABA}-n_{AB})}.\left(\frac{\boldsymbol{k}T}{\boldsymbol{h}\nu_A}\right)^{3(N_A-n_{AA})}.\left(\frac{\boldsymbol{k}T}{\boldsymbol{h}\nu_{BB}}\right)^{2(n_{BB}+n_{AB})}$$

$$\times\left(\frac{\boldsymbol{k}T}{\boldsymbol{h}\nu_{AA}}\right)^{\frac{1}{2}n_{AA}}.\left(\frac{\boldsymbol{k}^2T^2}{\boldsymbol{h}^2\nu_{ABA1}\nu_{ABA2}}\right)^{n_{ABA}}.\left(\frac{2\pi m_{BB}\boldsymbol{k}T}{\boldsymbol{h}^2}\delta_{BB}^2\right)^{\frac{1}{2}(n_{AB}+n_{BB})}$$

$$\times\left(\frac{2\pi m_{AA}\boldsymbol{k}T}{\boldsymbol{h}^2}\delta_{AA}^2\right)^{\frac{1}{2}n_{AA}}.\left(\frac{2\pi m_{ABA}\boldsymbol{k}T}{\boldsymbol{h}^2}\delta_{ABA}^2\right)^{\frac{1}{2}n_{ABA}}.$$

$$\exp\left\{-\frac{1}{\boldsymbol{k}T}[E_0+n_B\epsilon_B+n_A\epsilon_A+(n_{BB}+n_{AB})\epsilon_{BB}+n_{AA}\epsilon_{AA}+n_{ABA}\epsilon_{ABA}]\right\}$$

where $N=N_B+N_A+n_B+n_A$, E_0 = potential energy of crystal with no holes, provided the temperature is sufficiently high for specific quantal effects to be neglected. This is the case for temperatures at which the interdiffusion of metals is sufficiently rapid to be of interest. m_{AA} etc. are the reduced masses for the motion in the corresponding reaction coordinates.

For the equilibrium state F, and so $\ln F$, must be a maximum for variation of n_B, n_A, n_{BB}, n_{AA}, n_{ABA}, n_{AB}. The condition that this should be so leads to the equations

$$n_{AA}=(n_A-n_{AA}-n_{AB}-n_{ABA})\frac{\nu_A^3}{\nu_{AA}^2}n_{AA}e^{-\epsilon_{AA}/\boldsymbol{k}T},$$

$$n_{ABA}=(n_A-n_{AA}-n_{AB}-n_{ABA})\frac{4\nu_B^3}{\nu_{ABA1}\nu_{ABA2}}n_{ABA}e^{-\epsilon_{ABA}/\boldsymbol{k}T},$$

$$n_{AB}=(n_A-n_{AA}-n_{AB}-n_{ABA})\frac{7\nu_B^3}{\nu_{BB}^2}n_{BB}e^{-\epsilon_{BB}/\boldsymbol{k}T},$$

$$n_{BB}=(n_B-n_{BB})\ \frac{12\nu_B^3}{\nu_{BB}^2}e^{-\epsilon_{BB}/\boldsymbol{k}T}.\,n_{BB}.$$

Also

$$n_A/n_B=y/x\,.\,e^{(\epsilon_B-\epsilon_A)/\boldsymbol{k}T}\,.\,\frac{12N_A-n_A}{N_B-12N_A+n_A}=z$$

and

$$[N_B+N_A+(1+1/z)n_A]\,[N_B-12N_A+(1+1/z)n_A]=\frac{y}{z}n_A[N_B+(1+1/z)n_A]e^{\epsilon_B/\boldsymbol{k}T},$$

where

$$x=\frac{n_{\mathrm{A}}-n_{\mathrm{AA}}-n_{\mathrm{ABA}}-n_{\mathrm{AB}}}{n_{\mathrm{A}}}, \qquad y=\frac{n_{\mathrm{B}}-n_{\mathrm{BB}}}{n_{\mathrm{B}}}$$

and

$$u_j=\delta_j\sqrt{\frac{2\pi m_j}{\boldsymbol{k}T}}.$$

The solution of these equations is simple for $N_{\mathrm{B}} \gg N_{\mathrm{A}}$, $12N_{\mathrm{A}} \gg n_{\mathrm{A}}$, conditions corresponding to the initial physical assumptions. The velocity in the reaction coordinate j is

$$\sqrt{\frac{\boldsymbol{k}T}{2\pi m_j}};$$

so we have finally for the equilibrium state $n_{\mathrm{B}}=N_{\mathrm{B}}e^{-\epsilon_{\mathrm{B}}/\boldsymbol{k}T}$, $n_{\mathrm{A}}=12N_{\mathrm{A}}e^{-\epsilon_{\mathrm{A}}/\boldsymbol{k}T}$; and the numbers of jumps of different types taking place per second are

$$n'_{\mathrm{BB}}=n_{\mathrm{B}}\frac{12\nu_{\mathrm{B}}^3}{\nu_{\mathrm{BB}}^2}e^{-\epsilon_{\mathrm{BB}}/\boldsymbol{k}T}, \qquad n'_{\mathrm{ABA}}=n_{\mathrm{A}}\frac{4\nu_{\mathrm{B}}^3}{\nu_{\mathrm{ABA1}}\nu_{\mathrm{ABA2}}}e^{-\epsilon_{\mathrm{ABA}}/\boldsymbol{k}T},$$

$$n'_{\mathrm{AA}}=n_{\mathrm{A}}\frac{\nu_{\mathrm{A}}^3}{\nu_{\mathrm{AA}}^2}e^{-\epsilon_{\mathrm{AA}}/\boldsymbol{k}T}, \qquad n'_{\mathrm{AB}}=n_{\mathrm{A}}\frac{7\nu_{\mathrm{B}}^3}{\nu_{\mathrm{BB}}^2}e^{-\epsilon_{\mathrm{BB}}/\boldsymbol{k}T}.$$

Evidently, when a jump of the type AB takes place, the hole moves away from the A atom, whereas when a jump of type ABA takes place the hole moves round the A atom, remaining next to it. We are interested in the latter process as facilitating diffusion. Its relative probability,

$$\frac{n'_{\mathrm{ABA}}}{n'_{\mathrm{ABA}}+n'_{\mathrm{AB}}}=p,$$

must then be rather large if the effect is to be important. If this is so, we must also expect n'_{AA} to be very much larger than n'_{ABA}, so that when a hole arrives next to an A atom the latter will make many jumps between the two positions available for it before any other transition takes place.

Then it appears that any of eight equally likely ABA transitions may follow, since any of the four next neighbours to the two positions concerned may jump into either position. Since there is now no point in distinguishing the position of the A atom and that of the hole, we may speak of an AH pair and denote its position by the midpoint of the line joining the two neighbouring lattice points which constitute it. Then (figure 1) the positions available for the AH pair are the centres of the faces of the cells of a cubic lattice, the edge of whose unit cube is half that of the original atomic lattice. Each point in this lattice has eight nearest neighbours, and is the centre of symmetry of those eight. By a succession of ABA transitions, the AH pair may wander through the lattice, jumping from one point to one of its nearest neighbours.

Now since every point in this lattice is a centre of symmetry for its nearest neighbours, to any jump of the AH pair corresponds an equal and opposite jump which lands it on an equivalent lattice point. Thus to any given sequence of n jumps, represented by the ordered set of vectors $(r_1, r_2, \ldots, r_k, \ldots, r_l, \ldots, r_n)$ where every r_k has magnitude r equal to half the interatomic distance in the metal lattice, corresponds the whole set of 2^n equally likely excursions $(\pm r_1, \pm r_2, \ldots, \pm r_k, \ldots, \pm r_l, \ldots, \pm r_n)$. Now corresponding to any permutation of the signs of the other components, (r_k, r_l) may take the signs $(+,+)$, $(-,-)$, $(+,-)$, $(-,+)$. Thus if we take R^2, where $R=\Sigma r_k$, for each excursion, and sum over the whole set, the product terms such as $(r_k \,.\, r_l)$ cancel in groups of four and the sum is $2^n \,.\, nr^2$, so that $\overline{R^2}=nr^2$.

This gives the mean square distance travelled by an AH pair in consequence of n ABA transitions, since $\overline{R^2}$ is the same for all initial excursions $(r_1, \ldots, r_n)$ We require the mean square distance travelled by the A atom. This may initially have come from either of two positions distant r on opposite sides of the centre of the AH pair as first formed, and finally settles in either of two positions distant r on opposite sides of the final position of the centre of the AH pair. Thus, by the same argument as above, the mean square distance travelled by an A atom by the mechanism of AH pair formation, diffusion by n ABA jumps, AH pair dissociation by an AB jump, is $(n+2)r^2$. Now to a sequence of n ABA jumps preceding dissociation of the pair we must evidently assign a relative probability p^n. So the mean square distance travelled by an A atom each time a hole diffuses into a neighbouring position is

$$r^2 \frac{\sum\limits_0^\infty (n+2)p^n}{\sum\limits_0^\infty p^n} = r^2 \left[\frac{1+\sum\limits_0^\infty (1+n)p^n}{\sum\limits_0^\infty p^n}\right]$$

$$= r^2 \,.\, \frac{2-p}{1-p}.$$

The frequency with which an AH pair is formed must equal the frequency with which a pair dissociates. The frequency of dissociation of AH pairs over the whole system is n'_{AB}, so the frequency with which a given A atom enters into an AH pair is given by

$$n'_{\mathrm{AB}}/N_{\mathrm{A}} = 12e^{-\varepsilon_{\mathrm{A}}/kT} \,.\, \frac{7\nu_{\mathrm{B}}^3}{\nu_{\mathrm{BB}}^2}\, e^{-\varepsilon_{\mathrm{BB}}/kT}.$$

Thus the mean square distance through which an A atom diffuses in unit time by this mechanism is

$$r^2 \,.\, \frac{n'_{\mathrm{AB}}}{N_{\mathrm{A}}} \,.\, \frac{2-p}{1-p},$$

and the diffusion coefficient will be one-sixth of this; so

$$D = \frac{7}{2} a^2 \frac{\nu_{\mathrm{B}}^3}{\nu_{\mathrm{BB}}^2}\, e^{-(\varepsilon_{\mathrm{A}}+\varepsilon_{\mathrm{BB}})/kT} \,.\, \frac{\dfrac{4}{\nu_{\mathrm{ABA1}}\cdot\nu_{\mathrm{ABA2}}}\, e^{-\varepsilon_{\mathrm{ABA}}/kT} + \dfrac{14}{\nu_{\mathrm{BB}}^2}\, e^{-\varepsilon_{\mathrm{BB}}/kT}}{\dfrac{7}{\nu_{\mathrm{BB}}^2}\, e^{-\varepsilon_{\mathrm{BB}}/kT}},$$

where a = interatomic distance in lattice, and so

$$D \doteqdot 2a^2 \frac{\nu_{\mathrm{B}}^3}{\nu_{\mathrm{ABA1}}\cdot\nu_{\mathrm{ABA2}}}\, e^{-(\varepsilon_{\mathrm{A}}+\varepsilon_{\mathrm{ABA}})/kT}.$$

The measured activation energy for this process, $\epsilon_A+\epsilon_{ABA}$, may be very much less than that required for self-diffusion of B atoms, which is $\epsilon_B+\epsilon_{BB}$.

It may be observed that the mean square path of an A atom for a single collision with a hole is proportional to $e^{(\epsilon_{BB}-\epsilon_{ABA})/kT}$, so should increase rapidly with decreasing temperature. This might lead to apparent anomalies in diffusion through thin metal films at relatively low temperatures in cases where the mechanism of diffusion discussed above is important.

ACKNOWLEDGMENT

In conclusion, the author wishes to express his indebtedness to Mr. F. R. N. Nabarro and especially to Dr. F. C. Frank for fruitful discussions during the preparation of this paper.

REFERENCE

Johnson, 1939. *Phys. Rev.*, **56**, 814.

THE IMPERIAL COLLEGE HIGH-VOLTAGE GENERATOR

By W. B. MANN,*
National Research Council Laboratory, Chalk River, Ontario, Canada

AND

L. G. GRIMMETT, †
United Nations Educational, Scientific and Cultural Organization, Paris

MS. received 21 *October* 1946

ABSTRACT. The design and construction of two pressure-insulated electrostatic generators similar to those of Van de Graaff and Trump are briefly described. Voltage tests with one of the generators with mixtures of nitrogen and freon under pressure have shown it to be capable of producing voltages in excess of two million.

§ 1. INTRODUCTION

In the early summer of 1939 it was decided to instal a high-voltage electrostatic generator at the Imperial College to give around two million volts potential for the acceleration of positively ionized particles. The Medical Research Council had for some time also been considering a similar project for the provision of both high-voltage positive ions and electrons, for neutron, electron and x-ray investigations, but had been deterred from initiating such a programme on account of limited workshop facilities. It was therefore decided in the autumn of 1939 to make two such generators at the Imperial College to a common design following closely that of the pressure-insulated electrostatic generators of Van de Graaff and Trump. Work on both generators was started in the workshops of the Physics department at the Imperial College towards the end of 1939. In 1941, however, the war brought this work to a standstill. At this time the high-pressure tanks had been completed and a number of parts, such as the equipotential hoops and belt-guards, the insulators, upper electrode spinning and the charging-belt pulleys, had been designed and made. Such auxiliary equipment as the compressors, belts and the 15-h.p. motor for driving the charging belt had also been delivered.

---

* Formerly at Imperial College, London.

† Formerly at Radiotherapeutic Research Unit, Medical Research Council, Hammersmith Hospital, London.

In January, 1942, the Radiotherapeutic Research Unit moved to new quarters at Hammersmith Hospital, and it was found possible to resume work on the Medical Research Council's generator, the upper and lower charging-belt pulley supports, charging-belt motor support and generating voltmeter being designed and made; the design of these latter parts was chiefly the work of Mr. J. W. Boag and Mr. P. Howard Flanders, of the Radiotherapeutic Research Unit. Work on the Imperial College generator was resumed in the early summer of 1945, the designs for the upper and lower pulley supports, driving-motor support and generating voltmeter being adopted *in toto* from the Medical Research Council generator.

Both generators have now undergone voltage tests under pressure, and it is the purpose of this paper to give the results of the tests on the I.C. generator and a description of the generators in so far as they have followed a common design. Ion-source equipment for the upper electrode is under construction, but to different designs for each generator. Load tests will therefore be the subject of future and separate publications.

§ 2. GENERATOR DESIGN

A photograph of the Imperial College generator is shown in figure 1. The hoops are formed by rolling $\frac{3}{4}$-inch diameter tubing into rings 30 inches in diameter, each separated from the next by means of three textolite insulators. Each hoop is provided with four belt-guards similar to those fitted to the Van de Graaff and Trump generators, but they are so designed that their position can be adjusted after the column has been assembled. Bakelite spacer tubes are fitted over the belt-guards on every eighth hoop up the column. A photograph of one of the hoops and three textolite insulators is shown in figure 2. The upper three dozen hoops of the I.C. generator are rhodium plated, while the same number of the M.R.C. generator hoops are nickel plated to prevent corrosion. The lower hoops are polished prior to assembling.

The column is assembled on a flat base plate in order to have a minimum possible distance between the last accelerating electrode in the discharge tube and the target. The driving motor is therefore contained in a small pressure chamber supported on the under side of the main base plate, in the manner indicated in figure 3. This figure, which was completed in 1943, shows the lower pulley support only schematically as this part had not then been designed.

The upper electrode consists of an aluminium spinning and charge is conveyed to it by means of a Tilton endless-woven cotton belt. The upper belt pulley is supported on a fixed mount while the lower belt pulley is mounted on two cantilever supports which can be adjusted so as to take up any slack in the belt. By

Figure 2.

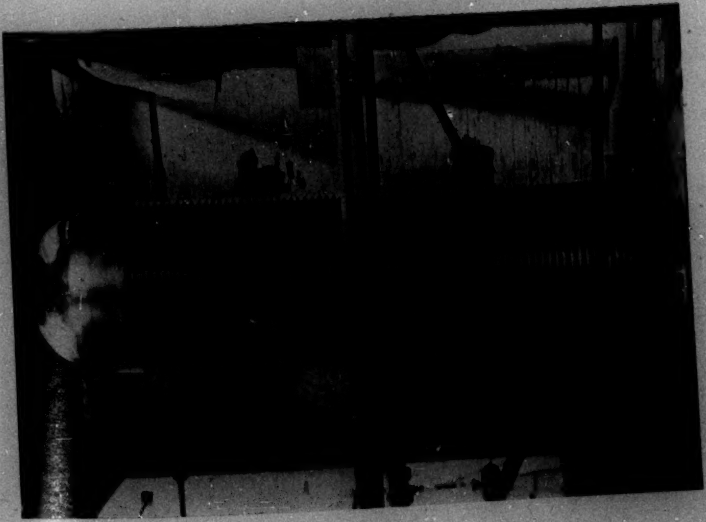


Figure 1.

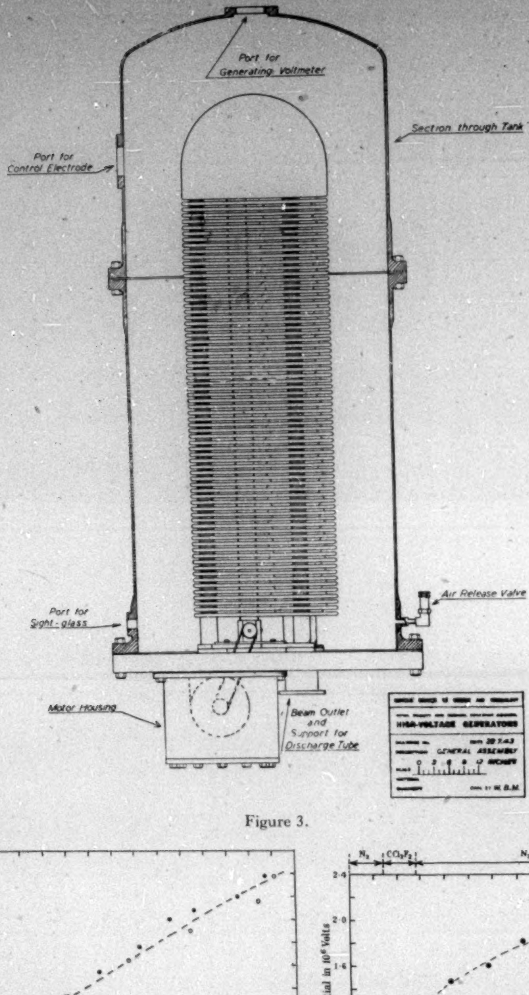
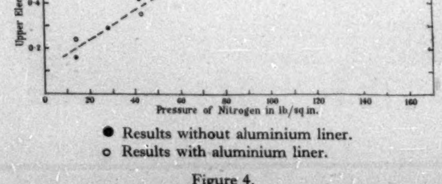

Port for
Generating Voltmeter
Section through Tank
Port for
Control Electrode
Air Release Valve
Port for
Sight-glass
Motor Housing
Beam Outlet
and
Support for
Discharge Tube
HIGH-VOLTAGE GENERATORS
GENERAL ASSEMBLY
0 2 4 6 8 12 INCHES
W.B.M.

Figure 3.



● Results without aluminium liner.
○ Results with aluminium liner.

Figure 4.

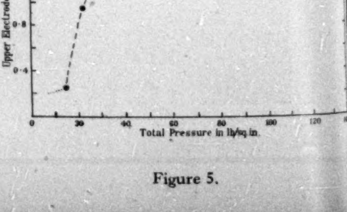


Figure 5.

differential adjustment, on either cantilever, the belt can be made to run in a central position on the pulleys. The fulcra of these cantilever arms can be seen in the foreground of the photograph of figure 1.

The column and electrode are some 9 feet high while the Tilton endless-woven belt is 19 feet long by 20 inches wide.

The uniform distribution of potential down the column is achieved by means of a small current drain from the top electrode. This can be realized either by means of corona gaps or resistors between adjacent hoops.

The former method has been employed in the I.C. generator and the latter for the M.R.C. generator. In the case of the I.C. generator, a gap was chosen to give, with negative point to plane, around 5 μa. drain at atmospheric pressure, i.e. with the upper electrode at a positive potential of about 300 kv. Resistors of 200 megohms each between adjacent hoops were employed in the M.R.C. generator. The corona gaps were constructed simply by soldering needles to paper clips, and can easily be slipped on or off one of the plane surfaces of a hoop, these surfaces consisting of $\frac{1}{32}$-inch thick brass sheet.

§ 3. VOLTAGE TESTS

In carrying out voltage tests, the generating voltmeter was used with a balancing voltage to give zero output, the balancing voltage at the null point being observed. This method will be described in a later publication.

Aluminium liners have been provided for the upper and lower pressure tanks for both the I.C. and M.R.C. generators. The I.C. generator was tested both without and with the upper liner, but not with the lower liner in position. The results of tests in compressed nitrogen both without and with the upper liner using the I.C. generator are shown in figure 4. The results obtained for the maximum voltage as a function of pressure are not sensibly different in the two cases. The only significant difference observed was that where sparks had been fairly generally distributed over the upper electrode without the liner they tended to terminate some six inches from the lower rim of the liner when it was installed. This result might indicate that it would be better to polish the steel tank itself or to continue the upper liner a foot or two into the lower section of the steel pressure tank. The charging currents (i.e. the lower spray comb currents) required at different pressures are shown in table 1. A short-circuit test of the current delivered to the upper electrode was also made for the highest pressure without the liner, a probe being inserted through the corona-control port of the upper section of the steel tank. By means of this probe the upper electrode was short-circuited through a milliammeter to earth. On increasing the charging current to 550 μa. the short-circuit current from the upper electrode increased to 450 μa. Thereafter the short-circuit current remained constant at 450 μa. with increase of the charging current to 1·5 ma. The theoretical saturation current for such a belt should be in the neighbourhood of 1·5 ma., and the discrepancy may be accounted for by leakage through the belt. It is to be hoped, therefore, that there will be an improvement in the current available for use in the discharge tube as the belt dries with use. The current loss due to corona is, as indicated by the figures of table 1, not large.

The voltage obtained at 1 atmosphere with nitrogen was observed always to be lower than that obtained with air at the same pressure. This is in agreement with

Table 1

Voltage test without liner			Voltage test with liner		
Pressure of nitrogen (lb./sq. in.)	Spray comb current (microamp.)	Voltage	**Pressure of nitrogen (lb./sq. in.)**	**Spray comb current (microamp.)**	Voltage
14	—	$0{\cdot}16\times10^6$	14	40	$0{\cdot}24\times10^6$
28	—	0·29	43	50	0·35
42	—	0·42	70	75	0·56
53	—	0·48	97	90	0·72
69	—	0·56	124	100	0·85
84	—	0·67	154	115	0·98
102	—	0·78	161	120	1·09
115		0·90			
126		0·94			
145		1·00			
157		1·09			

the fact that the dielectric strength of air is greater than that of nitrogen on account of the presence of negative oxygen ions.

Finally, measurements were made with the I.C. generator using mixtures of nitrogen and freon to give increased dielectric strength. The results obtained are shown in table 2 and figure 5.

Table 2

Pressure of nitrogen (lb./sq. in.)	Pressure of freon (lb./sq. in.)	Spray comb current (microamp.)	Voltage
14	8	100	$0{\cdot}95\times10^6$
14	12	120	1·12
14	14	100	1·18
27·5	14	120	1·47
44	14	140	1·61
58	14	150	1·82
72	14	100	1·75
84	14	120	1·96
96	14	120	2·03
108	14	160	2·20

The upper electrode, which was only resting in position, appeared to become unstable at the highest values of voltage and an intense rumbling noise developed. With the upper electrode securely fastened to the column, however, there appears to be no reason why the (voltage, pressure) curve should not be continued to around $2{\cdot}5\times10^6$ volts.

§ 4. CONCLUSION

At the time of writing, the ion source, accelerating tube and vacuum pumping systems are not yet finished. The voltage supply for the ion source inside the upper electrode is now complete but is, as mentioned previously, of a different design from that of the M.R.C. generator. Results for current output will therefore be the subject of future, separate, publications; this present paper deals only with those features of design which are common to both generators. Of the present authors, one (W. B. M.) has been associated with the project since its inception in 1939 till the present time, with an absence of some four years on war work; the other (L. G. G.) was associated with the work from its inception in 1939 until the autumn of 1944, besides being actively concerned for some time prior to 1939 with the question of providing such a high-voltage source for medica. work. In addition, Mr. J. W. Boag and Mr. P. Howard Flanders, of the Radiotherapeutic Research Unit, have been associated with the work on the M.R.C. generator since 1942.

§ 5. ACKNOWLEDGMENTS

Much of the work on the generators has been done in the Physics department workshop of the Imperial College, all but the heavy engineering parts of the I.C. generator having been made there and also the hoops and belt-guards, upper and lower pulleys and insulators of the M.R.C. generator. We are indebted to the following for their assistance and excellent work: Mr. W. Thompson, Mr. A. E. Davis, Mr. A. H. Bridger, Mr. F. Hart, Mr. L. Jenkins, Mr. S. Leach, Mr. W. Shand and Mr. W. S. Tonks. We are also indebted to Mr. C. W. Knighton and Mr. A. J. Campbell, of Messrs. Babcock and Wilcox Ltd., for the kind personal interest they took in the design and construction of the pressure tanks.

In the erection and testing of the I.C. generator one of us (W. B. M.) wishes gratefully to acknowledge the assistance which he has received from Dr. J. L. O. G. Michiels and the following research students and visitors: Mr. H. A. Dell, Mr. G. W. Green, Mr. G. A. Mann, Dr. P. Mottier, Mr. P. L. Parsons and Mr. G. C. Tavernier. Grateful acknowledgment is also made of the kind assistance and advice received from Mr. E. H. Laister, of Messrs. Johnson, Matthey and Co. Ltd., in connection with the rhodium plating of the brass hoops and from Mr. R. H. Walter and Mr. A. C. Whitney, of the Air Ministry, in connection with the installation of the high-pressure panel for handling nitrogen which was supplied by the Air Ministry in cylinders at 250 atmospheres. He also wishes to acknowledge with thanks a grant received from the Warren Fund of the Royal Society to cover the expenses of a trip to the United States in 1939. The cost of the I.C. generator has also been met by grants made from this fund.

The other author (L. G. G.) wishes to acknowledge with thanks the assistance rendered by Mr. J. W. Boag, Mr. P. Howard-Flanders, Mr. E. A. Rendle, and Mr. F. D. Pilling, of the Radiotherapeutic Research Unit.

We are both indebted to Professor Sir George Thomson for his interest in the work since its inception.

ON THE DETERMINATION OF ASPHERIC PROFILES

By E. WOLF and W. S. PREDDY,
University of Bristol

MS. received 1 *November* 1946

ABSTRACT. Exact parametric equations are deduced for the profiles of plano-aspheric lenses designed to produce axial stigmatism in a given axially symmetric pencil of rays. These formulae take an agreeably simple form in the special case where the point of stigmatism is at infinity. As an application, parametric equations are deduced for the corrector plate of the Schmidt Camera.

§ 1. INTRODUCTION

In many optical systems involving an aspheric surface, this surface, whose essential function is to improve the performance over the whole of a finite field, is designed so as to bring accurately to a focus the rays proceeding from the axial point of a selected object surface, which may be at infinity. Various exact and approximate formulae for the shape of such surfaces have been given in particular cases. In other cases, methods of successive approximation have been used, since a straightforward application of Snell's law leads to differential equations which are inconvenient to work with.

Using the principle of equal optical path, we deduce, with the help of a result proved in the next section, exact formulae for the surface-profile needed to annul the zonal aberrations of a given wave front. These formulae take an agreeably simple form in the special case where the point of stigmatism is at infinity. Although these formulae are exact, they involve an integral which in general has to be evaluated numerically.

As an application, parametric equations for the plate profile in the classical Schmidt Camera are obtained which have a wider range of validity than the formulae given by B. Strömgren (1935) and by J. G. Baker (1940).

§ 2. A PRELIMINARY RESULT

Let OH be the ordinate at a point O on the axis of symmetry of a system of rays proceeding from an axial source. We first derive an expression for the optical path difference between any two rays before reaching OH.

We denote by W a wave front corresponding to the system. Any surface which is orthogonal to all the rays is such a wave front, and the optical distance (O.D.) from the source to W is the same for every ray. Consider two neighbouring rays meeting W in A and C, and OH in B and D at heights h and $h+\delta h$ respectively. Let ω_1 denote the angle which the ray AB makes with the axis. Through B draw a line perpendicular to AB meeting CD in E. Finally let AB be denoted by I_h and CD by $I_{h+\delta h}$ (figure 1).

Then BE is a tangent to the wave surface passing through B, whence

$$I_{h+\delta h} - I_h = -\delta h \sin \omega_1 + O((\delta h)^2).$$

Dividing by δh, and proceeding to the limit as $\delta h \to 0$, we obtain

$$\frac{dI_h}{dh} = -\sin \omega_1,$$

whence

$$I_h = -\int^h \sin \omega_1 \, dh.$$

Let $I_{h_1}^{h_2}$ be the optical path difference between two rays which reach OH at heights h_2 and h_1.

Then

$$I_{h_1}^{h_2} = I_{h_2} - I_{h_1} = -\int_{h_1}^{h_2} \sin \omega_1 \, dh. \qquad \text{......(2.1)}$$

§ 3. PLANO-ASPHERIC LENSES

With the help of (2.1) we can determine the profile of the aspheric lens which will eliminate the zonal aberrations of a given incoming wave front. Only the case of aspheric lenses with one face plane is considered here, but it will be seen that the methods of §§ 3.1 and 4.1 (amounting essentially to the use of an ikonal) apply without change to the more general case where one face of the lens is of a given profile and the other face (the profile of which is to be determined to give axial stigmatism) is the last refracting surface of the system.*

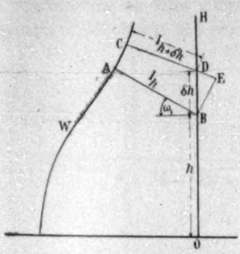


Figure 1.

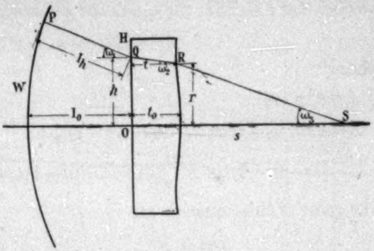


Figure 2.

§ 3.1. We first examine the mathematically simpler case where the aspheric surface of the lens faces the point of stigmatism. Figure 2 shows a ray, PQRS, proceeding from the wave front W towards the point of stigmatism S and meeting the refracting surfaces in Q and R. The origin O is taken at the point where the axis intersects the plane face of the lens.

D_h, the O.D. between P and S, has to be the same for every ray. We have

$$D_h = I_h + \frac{nt}{\cos \omega_2} + \frac{s-t}{\cos \omega_3},$$

and for the axial ray

$$D_0 = I_0 + nt_0 + s - t_0,$$

where t and t_0 are the thicknesses of the lens at R and O, s denotes† the distance

---

* This case has been treated also on the basis of the ikonal theory by Luneberg (1944). A different method, based on successive approximation, is described by Herzberger and Hoadley (1946); it is applicable also to the calculation of aspheric surfaces in the interior of a system.

† The case of a virtual image is also covered if s is allowed to take negative values.

OS, ω_1, ω_2 and ω_3 are the angles which PQ, QR, RS make with the axis (see figure 2), and n is the refractive index of the material of the plate.

Equating D_h to D_0, we obtain

$$I_0^h+\frac{nt}{\cos\omega_2}+\frac{s-t}{\cos\omega_3}-(n-1)t_0-s=0. \qquad \ldots\ldots(3.11)$$

From the figure

$$\cos\omega_3=\frac{s-t}{[(s-t)^2+(h-t\tan\omega_2)^2]^{1/2}}. \qquad \ldots\ldots(3.12)$$

Substituting for $\cos\omega_3$ in (3.11) we obtain

$$I_0^h+\frac{nt}{\cos\omega_2}+[(s-t)^2+(h-t\tan\omega_2)^2]^{1/2}-(n-1)t_0-s=0, \qquad \ldots\ldots(3.13)$$

where

$$\omega_2=\sin^{-1}\left(\frac{1}{n}\sin\omega\right). \qquad \ldots\ldots(3.14)$$

Rationalizing and rearranging (3.13) and substituting for I_0^h in terms of ω_2,* we find that t satisfies the quadratic equation

$$\left.\begin{aligned}
&At^2+2Bt\cos\omega_2+C\cos^2\omega_2=0,\\
&\text{where}\\
&A=n^2-1,\\
&B=h\sin\omega_2+s\cos\omega_2-n[s+(n-1)t_0+n\int_0^h\sin\omega_2\,dh]\\
&C=[2s+(n-1)t_0+n\int_0^h\sin\omega_2dh][(n-1)t_0+n\int_0^h\sin\omega_2\,dh]-h^2.
\end{aligned}\right\} \qquad \ldots\ldots(3.15)$$

The roots of this equation are

$$t_1=\frac{\cos\omega_2}{A}[-B+\Delta^{1/2}], \quad t_2=\frac{\cos\omega_2}{A}[-B-\Delta^{1/2}], \qquad \ldots\ldots(3.16)$$

where

$$\Delta=B^2-AC. \qquad \ldots\ldots(3.17)$$

It can easily be shown that only the solution $t=t_2$ satisfies the physical conditions.

From the figure, the radius r corresponding to h is given by

$$r=h-t\tan\omega_2, \qquad \ldots\ldots(3.18)$$

so that finally we have

$$t+ir=\frac{-e^{-i\omega_2}}{n^2-1}[B+\Delta^{1/2}]+ih. \qquad \ldots\ldots(3.19)$$

This equation gives the exact profile of the lens in terms of the parameter h.

§3.2. The case when the plane face of the lens is nearer to the point of stigmatism can be dealt with in a similar manner. Let O now be the point of intersection of the axis with the aspheric face of the lens and let t denote the thickness of the lens at Q (figure 3).

Then

$$D_h=I_h+\frac{t_0-t}{\cos\omega_1}+\frac{nt}{\cos\omega_2}+\frac{s-t_0}{\cos\omega_3}$$

and

$$D_0=I_0+nt_0+s-t_0.$$

---

* Not ω_1, as might appear more natural; the final solution takes a simpler form when expressed in terms of ω_2.

Equating D_h to D_0, we obtain

$$\frac{t_0-t}{\cos\omega_1}+\frac{nt}{\cos\omega_2}+\frac{s-t_0}{\cos\omega_3}-(n-1)t_0-s-\int_0^h \sin\omega_1\,dh=0 \quad \ldots\ldots(3.21)$$

From the figure

$$\tan\omega_3=\frac{h-(t_0-t)\tan\omega_1-t\tan\omega_2}{s-t_0}\;; \quad \ldots\ldots(3.22)$$

also

$$n\sin\omega_2=\sin\omega_3. \quad \ldots\ldots(3.23)$$

ω_2 and ω_3 can be eliminated between (3.21), (3.22) and (3.23), to give the equation for t; it is easier, however, substitute for ω_3 in terms of ω_2, insert values for the known quantities and solve the resulting equations by numerical methods.

The corresponding radius r is given by

$$r=h-(t_0-t)\tan\omega_1. \quad \ldots\ldots(3.24)$$

§4. PLANO-ASPHERIC LENSES: OBJECT OR IMAGE AT INFINITY

In the limiting case when $s\to\infty$, the lens converts into each other two ray-systems of which one is a parallel beam.

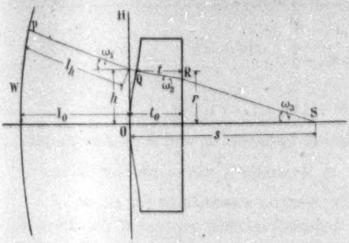


Figure 3.

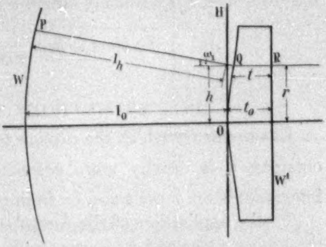


Figure 4.

§4.1. First we consider the case when the lens has its aspheric side towards the parallel beam. Dividing (3.15) by s and proceeding to the limit as $s\to\infty$, we obtain

$$2t\cos\omega_2(\cos\omega_2-n)+2\cos^2\omega_2[(n-1)t_0-I_0^h]=0$$

or

$$t=\frac{\cos\omega_2[(n-1)t_0-I_0^h]}{n-\cos\omega_2}, \quad \ldots\ldots(4.11)$$

and, as in §3.1,

$$r=h-t\tan\omega_2. \quad \ldots\ldots(4.12)$$

On substituting for I_0^h the complete solution can finally be expressed in the form

$$t+ir=\frac{e^{-i\omega_2}}{n-\cos\omega_2}\left[(n-1)t_0+n\int_0^h \sin\omega_2\,dh\right]+ih, \quad \ldots\ldots(4.13)$$

where

$$\omega_2=\sin^{-1}\left(\frac{1}{n}\sin\omega_1\right). \quad \ldots\ldots(4.14)$$

§4.2. The solution for the case where the lens has its plane side towards the parallel beam will now be obtained directly.

Since all the rays meet the plane face at right angles, it is a wave front (W′) of the system (figure 4).

The O.D. between the wave fronts W and W′ for the general ray is

$$D_h = I_h + \frac{t_0 - t}{\cos\omega_1} + nt$$

and for the axial ray is

$$D_0 = I_0 + nt_0,$$

whence

$$I_0^h + \frac{t_0 - t}{\cos\omega_1} + nt - nt_0 = 0,$$

giving

$$t = t_0 - \frac{I_0^h \cos\omega_1}{n\cos\omega_1 - 1}. \qquad \ldots\ldots(4.21)$$

We also have

$$r = h - (t_0 - t)\tan\omega_1. \qquad \ldots\ldots(4.22)$$

Substituting for I_0^h we finally obtain

$$t + ir = t_0 + ih + \frac{e^{i\omega_1}}{n\cos\omega_1 - 1}\int_0^h \sin\omega_1\, dh. \qquad \ldots\ldots(4.23)$$

§ 5. METHODS OF CALCULATION

In systems for which the explicit relation between ω_1 and h cannot be easily obtained, it is usually more practical to evaluate $\int_0^h \sin\omega_1\, dh$ by numerical integration from a ray trace, or from power-series expansions for $\sin\omega_1$.

By the application of the formulae deduced in this paper, the thickness of the aspheric lens and its corresponding radius can then be calculated for every value of the parameter h for which ω_1 and $\int_0^h \sin\omega_1\, dh$ have been determined. If more values are required they may be obtained by interpolation or curve fitting.

§ 6. APPLICATION: THE SCHMIDT CAMERA

We now apply equations (4.13) and (4.23) to find the profile of the figured surface of the Schmidt Camera. This system consists of a sperical mirror M and an aspheric plate P situated at its centre of curvature C. We first examine the arrangement (employed by Schmidt himself) in which the plane side of the plate faces the mirror (figure 5).

The rays which enter in a direction parallel to the axis pass through P, are reflected by M and focus at a point F near the paraxial focus of M. The figuring on the plate is such that it eliminates the axial spherical aberration of the system. We take the radius of the mirror as unity and denote CF by f. Further, we denote by δ the distance CO (measured as positive towards the mirror), O being, as in §4.1, the point of intersection of the plane face with the axis. Finally we take as parameter of the ray-system the angle ϕ between the axis and the line joining C with the point where the rays reach M.

From the geometry of the figure we find that

$$\omega_1=\tan^{-1}\frac{\sin\phi-f\sin 2\phi}{\cos\phi-f\cos 2\phi}, \qquad \ldots\ldots(6.1)$$

$$h=\frac{\sin\phi(f+\delta)-f\delta\sin 2\phi}{\cos\phi-f\cos 2\phi}, \qquad \ldots\ldots(6.2)$$

$$I_0^h=\frac{2\cos\phi-f\cos 2\phi-\delta}{\cos\phi-f\cos 2\phi}\cdot(1-2f\cos\phi+f^2)^{1/2}-(2-\delta-f). \qquad \ldots\ldots(6.3)$$

In substituting into (4.13) in terms of ϕ and equating real and imaginary parts, we obtain the following exact parametric equations for the plate profile

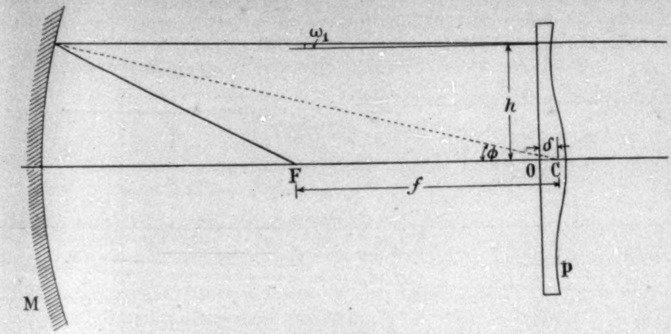


Figure 5.

if, as before, t denotes the thickness of the plate and r its corresponding zonal radius:

$$t=\frac{[n^2(1-2f\cos\phi+f^2)-\sin^2\phi(1-2f\cos\phi)^2]^{1/2}[(n-1)t_0-I_0^h]}{n^2(1-2f\cos\phi+f^2)^{1/2}-[n^2(1-2f\cos\phi+f^2)-\sin^2\phi(1-2f\cos\phi)^2]^{1/2}} \qquad \ldots\ldots(6.4)$$

and

$$r=\sin\phi\left\{\frac{f+\delta-2f\delta\cos\phi}{\cos\phi-f\cos 2\phi}-\frac{(1-2f\cos\phi)[(n-1)t_0-I_0^h]}{n^2(1-2f\cos\phi+f^2)^{1/2}-[n^2(1-2f\cos\phi+f^2)-\sin^2\phi(1-2f\cos\phi)^2]^{1/2}}\right\}, \qquad \ldots\ldots(6.5)$$

where I_0^h is given by (6.3).

The focal length CF $=f$ is usually chosen so that the chromatic aberration introduced by the plate is minimized. There are several different ways in which minimum chromatic aberration can be defined. B. Strömgren (1935) has shown that if h_n represents the height of the neutral zone (i.e. the radius of the zone for which a ray parallel to the axis passes through the plate undeviated)* and h_a the aperture radius, then to minimize (in Seidel approximation) the greatest angular departure from flatness over the whole plate, f should be chosen so that $h_n/h_a = 0{\cdot}866$. With this choice of f the maximum deviation of the ray parallel to the axis in the convergent sense is (again neglecting higher-order terms) the same as the maximum deviation in the divergent sense. Lucy (1940) showed that we minimize the integral of all deviations over the whole area of the aperture by giving the ratio h_n/h_a a value approximately equal to $0{\cdot}79$.

---

* In terms of h_n, $f=\frac{1}{2}(1-h_n{}^{-})^{1/2}$, as can be easily deduced from the figure.

To eliminate primary coma, the distance δ from C to the point O where the plane face meets the axis should be approximately equal * to t_0/n.

The camera may also be designed with the aspheric surface of the plate facing the mirror.† In this arrangement the aspheric surface passes through C (figure 6) so that ω_1, h and I_0^h are given by (6.1), (6.2) and (6.3) with $\delta = 0$. Then on substituting in (4.23) in terms of ϕ and equating real and imaginary parts we obtain the following equations for the plate profile, equivalent to (8) and (10) of Lucy's paper (1941):

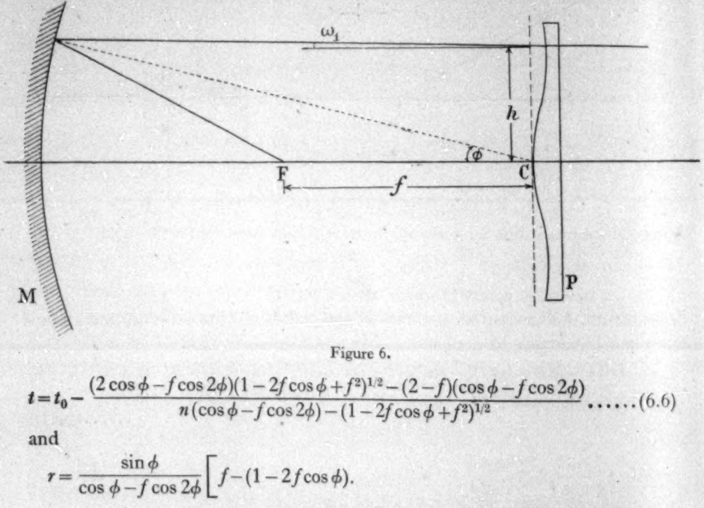


Figure 6.

$$t = t_0 - \frac{(2\cos\phi - f\cos 2\phi)(1 - 2f\cos\phi + f^2)^{1/2} - (2-f)(\cos\phi - f\cos 2\phi)}{n(\cos\phi - f\cos 2\phi) - (1 - 2f\cos\phi + f^2)^{1/2}} \quad \ldots\ldots(6.6)$$

and

$$r = \frac{\sin\phi}{\cos\phi - f\cos 2\phi}\left[f - (1 - 2f\cos\phi) . \frac{(2\cos\phi - f\cos 2\phi)(1 - 2f\cos\phi + f^2)^{1/2} - (2-f)(\cos\phi - f\cos 2\phi)}{n(\cos\phi - f\cos 2\phi) - (1 - 2f\cos\phi + f^2)^{1/2}}\right]. \quad \ldots\ldots(6.7)$$

For most astronomical purposes, the power-series expansions for the Schmidt Camera given by B. Strömgren (1935) and extended by J. G. Baker (1940) are sufficiently accurate and are simpler to work with than the formulae of this section. For sufficiently wide-aperture systems, however, these formulae are no longer

---

* In Lucy's paper the plate is incorrectly placed with its plane face passing through C (i.e. $\delta = 0$). Equations (6.4) and (6.5) become equivalent to Lucy's results if δ is given this value.

† A procedure apparently adopted by some authors for the sake of mathematical convenience. In the astronomical case it has the effect of producing sharp " ghosts " unless the plate is given a slight overall " bending ". Such a bending has little effect on the image errors, but it complicates the practical construction of the plate.

adequate. As J. G. Baker (1940) remarks: "for such extreme cases either a differential correction or else an integration based on ray-tracing formulae must be carried through". In the present section we have provided formulae suited to such extreme cases.

ACKNOWLEDGMENT

In conclusion we should like to express our thanks to Dr. E. H. Linfoot for much valuable assistance with the preparation of this paper.

REFERENCES

Baker, J. G., 1940. *Proc. Amer. Phil. Soc.*, **82,** 323.
Carathéodory, C., 1940. *Hamburg. Math. Einzelschr.*, **28.**
Glancy, A. E., 1946. *J. Opt. Soc. Amer.*, **36,** 416.
Herzberger, M. and Hoadley, H. O., 1946. *J. Opt. Soc. Amer.*, **36,** 334.
Lucy, F. A., 1940. *J. Opt. Soc. Amer.*, **30,** 251 ; 1941. *Ibid.*, **31,** 358.
Luneberg, R. K., 1944. *Mathematical Theory of Optics* (Brown University).
Strömgren, B., 1935. *Vierteljahreschrift der Astronom. Ges.*, **70,** 65.

CORRIGENDA

"The fundamental concepts concerning surface tension and capillarity", by R. C. Brown (*Proc. Phys. Soc.*, **59**, 429 (1947)).

Page 436, equation (4): insert minus sign in front of "$p_s t$".
Page 445, equation (12): for "$-\gamma_{S(S)}$" read "$-\gamma_{L(S)}$".
Page 445, beginning of seventh line below the figures: for "$-\gamma_{L(L)}$" read "$\gamma_{L(S)}$".

"The short-period time-variation of the luminescence of a zinc sulphide phosphor under ultra-violet excitation", by Mary P. Lord, A. L. G. Rees and M. E. Wise (*Proc. Phys. Soc.*, **59**, 473 (1947)).

Page 473, insert "Material Research Laboratory" before "Philips' Lamps Ltd".
Page 477, line 3, insert after "present". "The phosphor also contained 0·48% magnesium, which does not cause activation." (The authors thank Mr. C. G. A. Hill for this information.)

REVIEWS OF BOOKS

Theory and Application of Mathieu Functions, by N. W. McLachlan. Pp. xii + 401. (London: Geoffrey Cumberlege, at the Oxford University Press, 1947.) 42*s*.

Mathieu functions satisfy the differential equation

$$\frac{d^2y}{dz^2}+(a-2q\cos 2z)y=0, \qquad \ldots\ldots(1)$$

just as Bessel functions satisfy

$$\frac{d^2y}{dz^2}+\frac{1}{z}\frac{dy}{dz}+\left(1-\frac{n^2}{z^2}\right)y=0.$$

In both instances the equation defines y as a function of z and of certain parameters (a and q for the former, n for the latter), and we have to include the case where z is pure imaginary. For both equations we then write the solution as a function of z/i and call it a *modified* function.

In the case of Bessel equation, there can only be two independent solutions, and the general solution is a linear combination of these. Yet we are familiar with the fact that particular combinations are of such frequent occurrence, and such general usefulness, that we treat them as independent solutions of equal standing, so that we have not only J_n and Y_n, but also *ber* and *bei*, *ker* and *kei*, $H_n^{(1)}$ and $H_n^{(2)}$, K_n and N_n as well as I_n, and we expect a reader to be familiar with all of them.

The Bessel functions are much better known than those of Mathieu, both in the sense that many people are familiar with the properties of them, and also in the sense that more of their properties have been sought out and placed on record in mathematical literature. This arises from a combination of many causes, of which the greater intrinsic complexity of the study of Mathieu functions, and the fact that applied mathematicians had not called urgently for them, are doubtless important. The instrinsic difficulty arises in part from the fact that there are two parameters, and in part from the fact that there is a trigonometric instead of an algebraic term in the differential equation. Applied mathematicians did not ask for them, they would say, because their investigations did not lead that way; but we may suspect that they tended to avoid investigation which would have called for Mathieu functions, just because their properties were not fully catalogued and particularly because there were no tables of them—a fact which in its turn may be traced back, at least in part, to the existence of the multiplicity of parameters.

Nevertheless, there is a literature, and Dr. McLachlan has performed a really stupendous task in sorting out and presenting afresh, and in a co-ordinated manner, the whole theory of the functions. Naturally, he has had to fill in many of the gaps, and has done so with a modesty which makes it difficult to pick out for mention his own original contributions.

As with the Bessel functions, it is found desirable to define and work with a number of different fundamental pairs of solutions, and these are here systematized under a notation which makes it fairly easy to keep their special peculiarities in mind. These functions, however, possess one property which has no analogue in Bessel functions, though it has analogies in other equations familiar to the physicist. If we seek a solution of (1) which shall be periodic in z (as occurs naturally, if z is an angle), we find that there must be a relation between a and q. In other words, there are then *characteristic values* of a, imposed by the very nature of the solution. This is a situation familiar in wave mechanics, where the requirement that a solution of Schrödinger's equation shall be regular at infinity imposes characteristic values on the energy. Dr. McLachlan gives separate consideration to solutions of this periodic type, and puts workers much more deeply in his debt by considering in adequate detail the problem of numerical calculation of the solutions and of the characteristic numbers.

Two of the most useful types of solution appear to be those obtained by expansion in terms of trigonometric functions on the one hand and in terms of Bessel functions on the other. The general properties of orthogonality, so important for fitting solutions to boundary conditions, are treated, and asymptotic expansions are fully dealt with. A most valuable appendix, due to the late Prof. Ince, gives a table of the characteristic numbers. Ince, in fact, did a great deal of work on these functions, and published relatively extensive tables which are not reproduced in this book, though references are supplied.

Not only has Dr. McLachlan given us the whole corpus of useful recorded knowledge of these functions, but he discusses also the physical problems in which they play a prominent part. One such problem arises when the wave equation in elliptic co-ordinates is *separated*. For the one variable, the result is an ordinary Mathieu equation, and for the other a *modified* one, and it was in this connexion that Mathieu was first led to study the functions. Lunar theory leads to a differential equation which is a generalization of Mathieu's, but the theory of the two is very similar, and MacLachlan includes a chapter on the subject. Other physical applications have mostly arisen in recent years—Oseen's hydrodynamic equation, the theory of frequency modulation and of wave guides—but a striking exception is the detailed explanation of Melde's experiment with a tuning fork, where the string maintained in oscillation by it has a frequency which is a sub-multple of that of the fork.

It is clear that the contemporary development of physics will cause a demand for these functions. It is at least probable that this book will encourage the examination of phenomena in which these functions are involved, and which, without it, would have been set aside for future examination.

J. H. A.